Utrecht University

Student number: 6866409

Unlocking the potential of bootstrapping:

A journey towards balanced and reliable synthetic data

A framework for evaluating Bootstrap in the context of synthetic data generation

Master Thesis
Business Informatics, Faculty of Science

Utrecht University

Graduate School of Natural Sciences

Information Science (MSc)

Authored by

Denisa-Loredana Caragea

Utrecht, *14 - November - 2023*

Thesis Advisor: Dr. Gerard Wagenaar — Department of Computer Science

Second Reader: Dr. Ioana Karnstedt-Hulpus — Department of Information and Computing Science

Disclosures:

- I affirm that I have written the dissertation myself and have not used any sources and aids other than those indicated.
- I affirm that I have not included data generated in one of my laboratory rotations and already presented in the respective laboratory report.

Date / Signature:

Abstract

Synthetic data generation is an essential technique in data analysis and machine learning, playing a crucial role in complementing existing data sets and addressing the various challenges associated with their analysis. Synthetic data have significant utility where original data sets are limited, inaccessible, or insufficiently diverse. By incorporating synthetic data, it becomes feasible to augment the dataset size, thereby facilitating the effective implementation of diverse analysis and machine learning algorithms. However, generating synthetic data does not come without challenges and risks. Among the most significant challenges are class imbalances in the datasets, where certain classes are under-represented, which can affect the results and correct interpretation of the analysis. In addition, data confidentiality must be maintained, especially for datasets containing sensitive information.

This research addresses these challenges by focusing on evaluating a synthetic data generation method based on Bootstrap resampling. Inspired by Bootstrap, this paper proposes the "Fusionstrap" framework. This framework integrates the stratified Bootstrap method with sample post-processing techniques to address class imbalances in datasets, enhance the diversity and accuracy of synthetic data, and concurrently uphold the levels of usefulness and confidentiality. The effectiveness of this approach is assessed through an experimental case study, where synthetic data is generated, and the performance of our proposed framework is analyzed in comparison to the basic CTGAN and Synthpop methods using three datasets. The training data was collected and preprocessed using appropriate tools and techniques. Our evaluation metrics capture improvements in synthetic data quality and provide detailed insight into the strengths and weaknesses of the evaluated methods. We conclude that the application of the "Fusionstrap" framework aspires to generate accurate, balanced and representative synthetic data. Furthermore, it could be used as an aid to data generation to improve accuracy in the case of an unbalanced data set.

Keywords:
Synthetic Data, Preprocessed Techniques, Stratified Bootstrap, Class Imbalances, Post-processing Techniques, Utility, Privacy

Table of Contents

Chapter 1
Introduction

In today's age of technology, we collect a huge amount of data that can support decision-making in a variety of fields. This use of data may affect society directly or indirectly. For example, in healthcare, data collected from clinical trials and medical record systems can be used to identify relevant trends and patterns in the evolution of diseases and to support medical decisions [1]. At the same time, census data can provide a wide range of demographic, social, economic and cultural information about the population. These data are of major importance for public policy formulation, strategic planning, private sector decision-making and understanding social change [2].

In many situations, however, the data collected may present certain challenges, such as class imbalances [3]. This means that some classes of data occur more frequently than others, which can lead to inadequate learning of analytical models.

Classification problems, such as class imbalance, can have a significant impact on decisions made based on medical data in the following aspects [4]:

- Misdiagnosis: Class imbalance can lead to underrepresentation of certain rare or unusual medical conditions in the dataset. This can lead to incorrect or delayed diagnosis of these conditions, which can adversely affect the treatment and prognosis of affected patients.
- Inadequate treatment: If certain groups of patients with specific conditions are underrepresented in medical data, the effectiveness of certain treatments may be underestimated or overestimated. This can lead to the administration of inappropriate treatments for patients with specific medical needs.
- Risk assessment and prognosis: Class imbalance can affect the risk assessment and prognosis of patients. Underrepresented groups may have underestimated or neglected risks or prognoses, which may lead to inadequate management of their health status.
- Personalization of healthcare: To personalize treatment and healthcare, it is essential to fully understand the individual needs of patients. Class imbalance can affect this understanding and prevent identification of the specific needs of some patient groups.
- Medical research: Class imbalance can influence the results of clinical trials and epidemiological analyses, leading to inappropriate generalization of results and their application to the entire population.

Also, class imbalance can affect decisions made on the basis of census data in several ways [5]:

- Bias in Public Policy: If certain groups or population categories are underrepresented in census data, policy decisions may not take into account the needs and interests of these groups, leading to inappropriate and unfair policies.
- Inaccuracy in estimates: Class imbalance can lead to an incorrect representation of population distribution, which can affect demographic and economic estimates and projections, making them less accurate.
- Underestimation of social needs: If certain population categories, such as minorities or marginalized groups, are underrepresented in census data, their specific problems and needs may be underestimated or neglected in the decision-making process.
- Limiting social analysis: Class imbalance can affect social analysis and research, making it less representative and less relevant to understanding the complexity of society.

To address these issues, it is essential to adopt appropriate methods for handling class imbalance in datasets. Researchers and practitioners have developed various methods and techniques to address this asymmetry, which may include rebalancing the dataset through oversampling or undersampling, model-based techniques, and generating synthetic data [3]. Generating synthetic data is a promising approach in managing class imbalance as it can help provide a balanced and more representative data set. By generating synthetic examples that faithfully capture the features of the initial data, this methodology can enhance the efficacy of machine learning models, ensuring improved generalization across diverse domains

such as medicine, finance, marketing, and beyond. For example, in medical classification problems, synthetic data can be created to ensure a balance between rare and common classes [6].
Once the context of the research is defined, this chapter continues with the problem and motivation in Sec. 1.1. This is followed by the main research questions addressed in Sec. 1.2 together with the contributions of this thesis in Sec. 1.3. Finally, Sec. 1.4 concludes with an outline of how this research is organized.

1.1 Problem and motivation

In today's context, data has become an essential element in decision-making and the development of artificial intelligence algorithms. Medical data and the census are two extremely important fields where managing class imbalance in data sets can significantly influence decisions and outcomes. In the field of health, data collected from clinical trials and medical record systems can be used to identify relevant trends and patterns in disease progression and support medical decisions. Take, for example, a medical classification problem in which it is desired to identify a patient's risk of developing a serious condition such as diabetes. In such a scenario, class imbalance can affect how the classification algorithm builds its models and lead to an unfair representation of different patient groups. Underrepresented classes, such as patients at low risk of diabetes, may be inadvertently overlooked, leading to underestimation of the true risk. On the other hand, the classifier may be over-represented for major classes, which may lead to incorrect and unfair results for patients in minor groups. In the context of the census, the data collected are essential to understand the demographic composition of a country or region and to make decisions about public policy, resource allocation and infrastructure development. Take, for example, a recent census in a country where it is desired to classify the population into various socio-economic groups such as education level, occupation and income. In such a scenario, class imbalance may affect the accuracy and representativeness of census results. Minority groups, such as those with low incomes or low levels of education, may be underrepresented in census data, leading to an incomplete understanding of the situation of these vulnerable populations. On the other hand, major groups, such as those with high incomes or high levels of education, may be over-represented, which can lead to a distorted picture of the true distribution of the population.
One possible approach to solving this problem is to equalize the imbalance in the underlying data set by generating enough examples for the underrepresented classes. Generating synthetic data is a solution for solving class imbalances in data sets, but the practice faces certain challenges: it is difficult to reproduce complex characteristics of real data; there is the possibility of losing some functions or characteristics necessary for the replication process; the flexible nature of synthetic data makes it biased in behavior, etc. One barrier to using synthetic data for real-world analysis is uncertainty about its usefulness and confidentiality. For synthetic data to be beneficial, it must yield valid results in statistical analyses. Surprisingly, there has been a relative lack of research into measuring the usefulness of synthetic data. Also, synthetic data cannot be guaranteed to ensure the confidentiality of all records from the original data [9].
The literature analyzes different algorithms and tools with which synthetic data can be generated. For example, Monte Carlo simulation can be useful when real data are available to be simulated and the distribution parameters are known [10]. An alternative for producing synthetic data involves employing deep learning models, such as generative adversarial networks (GANs). This becomes especially valuable when a significant volume of data is required, and the underlying distributions are not adequately known. Other examples would be decision trees, reverse engineering techniques, and iterative proportional fitting. Most synthetic data generators [11] require a lot of user specification or knowledge of the underlying data distribution to be synthesized. Furthermore, these approaches [11] do not guarantee that the resulting data sets provide the desired data distribution and correlations between attributes. According to the theory, when specific conditions are met, such as having a large sample size, the sampling distribution tends to approximate a normal distribution, with the standard deviation of the distribution equating to the standard error. However, if the sample size is insufficient or when the assumption of a normal sampling distribution is not valid, determining the standard error of the estimate becomes challenging. Consequently, drawing meaningful conclusions from the data becomes more complex in such situations. Bootstrapping emerges as a valuable solution to tackle the aforementioned challenges, particularly in

scenarios where the sampled population is intricate or unknown, or when obtaining the desired distribution of sampling statistics proves challenging. In cases where the population is identified, repeated sampling becomes a viable approach to characterize the desired sampling distribution, subject to Monte Carlo error. Each generated data set possesses a distinct set of sample statistics, encompassing measures such as mean, median, and standard deviation. In bootstrapping procedures, the distribution of these sample statistics across the simulated samples is employed as the sampling distribution. The mentioned distribution is applicable for computing precise confidence intervals and conducting pertinent hypothesis tests within the realm of bootstrapping. Bootstrap intervals and p-values can be regarded as real-world approximations. Although other statistical techniques used to determine confidence intervals require knowledge of the mean or standard deviation for the selected population, bootstrapping requires nothing more than the sample [12]. However, generating synthetic data by the Bootstrap method may have some problems, such as lack of diversity and precision. In this regard, post-processing techniques such as filtering or removing outliers are needed to improve the quality of the synthetic data. In addition, combining multiple Bootstrap datasets can lead to greater diversity and greater accuracy of synthetic data.

In the context of this thesis project, we explore two sets of techniques for generating synthetic data for comparison purposes. The first approach involves the utilization of deep learning methods, such as CTGAN (Conditional Tabular GAN), which is a variant of Generative Adversarial Networks (GANs) specifically designed for tabular data [7]. GANs are machine learning algorithms that consist of two competing neural networks: the generator and the discriminator. The generator is tasked with acquiring the skill of generating synthetic data, while the discriminator is focused on developing the ability to differentiate between real and synthetic data. Throughout the training process, these two networks engage in a zero-sum game. The generator endeavors to produce synthetic data with heightened realism to deceive the discriminator, while the discriminator strives to enhance its proficiency in distinguishing between real and synthetic data. The advantage of CTGAN is that it improves the ability of GANs to generate tabular data while maintaining the structure and complex features of the original datasets. By training on real data and then generating synthetic data based on it, CTGAN can create new examples that preserve the original distribution and data patterns.

The second approach - Synthpop is a statistics-based method and relies on refolding techniques [8]. Essentially, Synthpop generates synthetic data by replicating the distributions and structure of the original dataset. This method is useful for datasets with complex and dependent variables. Synthpop is based on an iterative refitting process, where the distributions of the variables are adapted and adjusted according to the original data. Thus, the created synthetic datasets retain more complex correlations and dependencies from the original data, making them more realistic and representative.

To tackle these challenges, we chose to adopt a hybrid approach called "Fusionstrap", which combines Stratified Bootstrap [32] with the Gaussian Copula Synthesizer [45]. This choice is based on the advantages offered by Stratified Bootstrap in addressing class imbalance, along with the ability of the Gaussian Copula Synthesizer to capture complex correlations between variables. Comparisons with established methods such as CTGAN and Synthpop can highlight the merits and significance of "Fusionstrap" in class balancing and enhancing the efficacy of machine learning models.

In short, the motivation of this research is given by the following considerations:

- Stratified Bootstrap is a solution for solving class imbalances in datasets;
- Most synthetic data generators require user specification and knowledge of the data distribution, while the Bootstrap only requires the sample data;
- Continuous variables exhibit highly skewed distributions that are difficult to model and reproduce authentically. Bootstrapping may encounter difficulties in situations where the underlying population is intricate or unfamiliar, or when obtaining the desired distribution of sampling statistics proves challenging;
- Bootstrap can be effectively applied to any type of variable, whether numerical or categorical;
- Bootstrap is widely recognized in statistical theory, but is quite underused in practice, although it fits well with the computer age. Although datasets created with bootstrap procedures are successfully used in many applications, the scientific community's understanding of the power of bootstrapping still

remains unknown. For this reason, the problem of choosing a reliable bootstrap procedure for the domain of synthetic data generation remains open.

The primary objective of this study is to assess the effectiveness of the "Fusionstrap" approach in producing synthetic data. Additionally, the research aims to investigate strategies for enhancing the quality of this data through post-processing techniques, including filtering, outlier removal, and the amalgamation of synthetic datasets obtained through Stratified Bootstrap. Simultaneously, we will prioritize safeguarding the confidentiality of the initial dataset. A particularly important aspect is the evaluation of the data synthesized with "Fusionstrap" through a rigorous comparison with the CTGAN and Synthpop methods.

1.2 Research Questions

The primary objective of this thesis is to create and assess a tool proficient in generating synthetic data.

Thus, the research will focus on the following:

MRQ: **Can "Fusionstrap" improve the quality of synthetic data over known methods such as CTGAN or SYNTHPOP?**

To address the central research question, specific sub-questions were formulated to delineate the knowledge to be acquired and the requisite research activities.

RQ1: To what extent can „Fusionstrap" ensure the utility of the data generated?

RQ1.1 How can the utility of synthetic data be measured?

RQ1.2 What is the utility level of the "Fusionstrap" framework compared to other data synthesis methods (CT GAN and SHYNTPOP)?

RQ1.3 To what extent does "Fusionstrap" resolve class imbalances compared to CTGAN and SYNTHPOP?

RQ2: To what extent can "Fusionstrap" ensure the confidentiality of the original data?

RQ2.1 How to quantify the disclosure risk of synthetic data?

RQ2.2 How well does "Fusion strap" protect the confidentiality of the original data compared to other methods (CTGAN and SYNTHPOP)?

1.3 Expected Contributions

Concerning the anticipated contributions of this investigation to the domain of synthetic data generation, encompassing both practical and scientific perspectives, they comprise:

Practical Contribution:

- Development of the "Fusionstrap" method: This adaptive method uses the Stratified Bootstrap to resolve class imbalances in datasets, especially complex and dependent ones such as medical and census data. "Fusionstrap" integrates Boostrap sample post-processing techniques to improve the accuracy and variability of synthetic data;
- Generating balanced synthetic data: Using the "Fusionstrap" method, balanced synthetic data sets can be generated that preserve the correct proportion between under-represented and over-represented classes. This generated synthetic data holds value in enhancing the performance of machine learning models, particularly in scenarios where class imbalance adversely impacts accuracy and the relevance of decision-making;
- Evaluation of the utility and confidentiality of synthetic data: In the thesis, a rigorous evaluation of the utility and confidentiality of the synthetic data generated by the "Fusionstrap" method was carried out, comparing them with those generated by the CTGAN and Synthpop methods. This evaluation offers a more profound comprehension of the quality of synthetic data and its possible applications in research and decision-making.

Scientific Contribution:

- Solving the problem of class imbalance: The thesis proposes an efficient approach to balance the distribution of classes in data sets, so that machine learning models are fairer and more accurate in decisions.
- Exploration and comparison of synthetic data generation methods: By carrying out a detailed experiment and comparing the performances of the "Fusionstrap" method with those of the CTGAN and Synthpop methods, the thesis makes a scientific contribution to the wider evaluation of various synthetic data generation techniques and the identification of advantages and their limitations.
- Improving the quality of synthetic data: The study contributes to a deeper understanding of the quality of synthetic data generated by different methods. By identifying existing challenges and limitations, the thesis can pave the way for new research and development to improve the quality and use of synthetic data in diverse fields.
- Potential applications in data-based research and decision-making: By validating the "Fusionstrap" method and demonstrating the utility and confidentiality of the generated synthetic data, the thesis contributes to the development of a practical and beneficial tool in data-based research and decision-making.

In summary, since Bootstrap is a widely used method in data analysis and generating statistical estimates, understanding how an appropriate Bootstrap method can be used to generate synthetic data could make a significant contribution to the research field and improve how synthetic data is used and understood in data analysis.

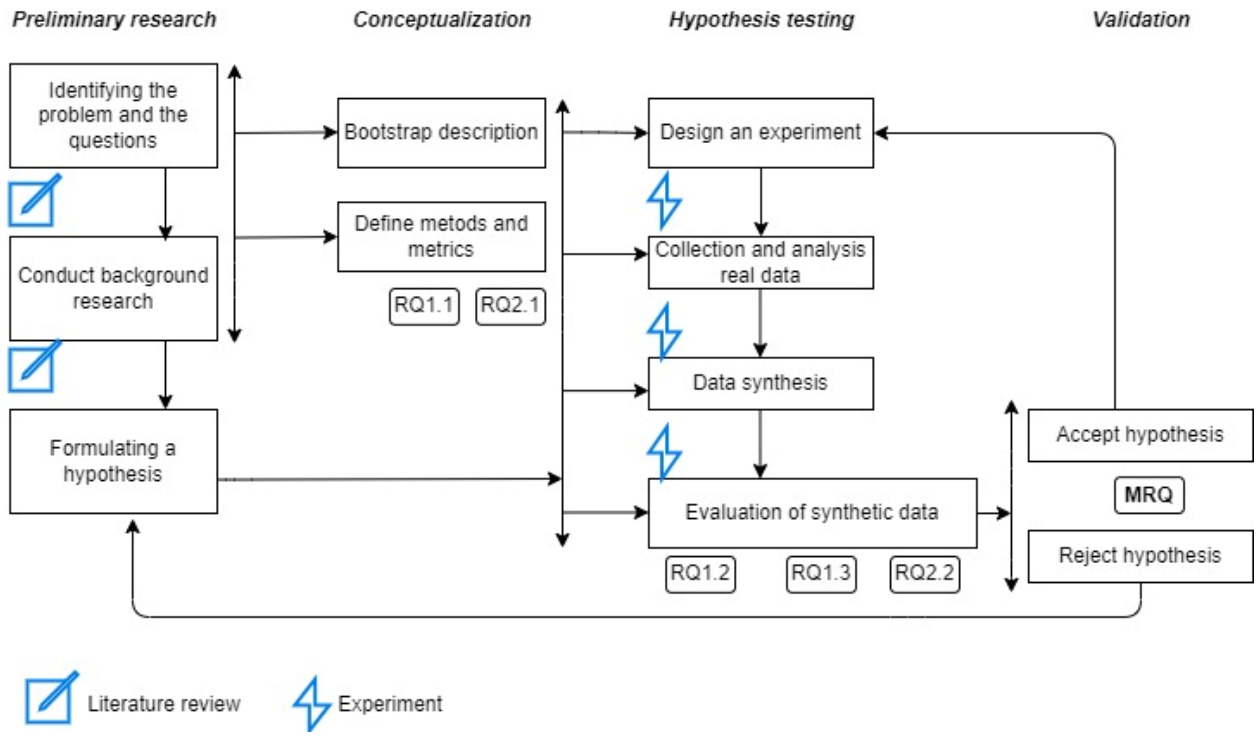
1.4 Outline of Thesis

The rest of the report delves into the justification and implementation of the research. Chapter 2 elucidates the research methods, providing an in-depth explanation. Chapter 3 covers relevant preliminary work on the concepts and notions used and explains the Bootstrap resampling process. Chapter 4 presents the proposed "Fusionstrap" framework as a hybrid method for synthetic data generation. In Chapter 5, the experimental design is presented, and the data sets used for the experiments are described, and the results are presented in Chapter 6. Finally, Chapter 7 summarizes the conclusions of this thesis by reviewing the research questions, identifies the limitations of "Fusionstrap" and defines research avenues for future work.

Chapter 2
Research approach

Given that the purpose of the research is to conduct an experiment to test the hypothesis that "Fusionstrap" could generate synthetic data of better or at least comparable quality to known methods such as CTGAN or SYNTHPOP, this thesis will use a scientific approach in methodological research [13]. The research framework for this study is presented in Figure 1 and presents the steps that will be followed to achieve the research goal (sec.1.2). In section 2.1, the research questions are justified, and the research methods used to answer these questions are linked. Section 2.2 offers a detailed description of the research methods, while Section 2.3 concludes by examining potential threats to the validity of the research.

Figure 1
Research steps



The first step, preliminary research, will consist of identifying the problem, formulating research questions, and establishing a hypothesis to be tested. The conceptualization phase (the next step of the research methodology) will be based on the theory of synthetic data generation and evaluation, as well as the theory of the Bootstrap method. In this second step, preliminary ideas will be explored and the tools, methods, and metrics that will be used to test the hypothesis will be described. The results of these first steps will provide answers to research questions RQ1.1 and RQ2.1. The third step, hypothesis testing, involves the description of the experimental framework, the collection and analysis of original data, the generation of synthetic data and their evaluation. The results will provide answers to research questions RQ1.2, RQ1.3 and RQ2.2. Finally, the validation phase will lead to either accepting the hypothesis or rejecting it. Accepting the hypothesis involves repeating the experiment on various types of data to conclude whether the hypothesis can be generalized. If the hypothesis is rejected, it can be modified and retested until it is consistent with observed phenomena and test results.

2.1 Justification of research questions

In this section, we will explore and justify each of the research questions we have formulated previously. Through analysis and argumentation, we will demonstrate the importance of these questions in our research and how they contribute to the achievement of the proposed objectives.

RQ1: To what extent can „Fusionstrap" ensure the utility of the data generated?

RQ1.1 How can the usefulness of synthetic data be measured?

RQ1.2 What is the utility level of the "Fusionstrap" framework compared to other data synthesis methods (CT GAN and SHYNTPOP)?

RQ1.3 To what extent does "Fusionstrap" resolve class imbalances compared to CTGAN and SYNTHPOP?

The performance of the "Fusionstrap" method in generating synthetic data is an important concern for the development of effective solutions in handling class imbalance and improving the performance of machine learning models. To assess and verify the efficacy of this approach, it is crucial to examine the extent to which "Fusionstrap" can produce synthetic data of high quality while retaining the fundamental characteristics inherent in the original data. To evaluate the performance of the "Fusionstrap" method in generating synthetic data, we will perform an extended experiment on three different datasets. We will apply the "Fusionstrap" method for generating synthetic data on each dataset and evaluate the quality of the generated data.

RQ1.1: Assessing the utility of synthetic data is pivotal for gauging their effectiveness as substitutes for real data in practical applications. As the employment of synthetic data has implications for the outcomes of machine learning models and decision-making processes, it becomes crucial to identify and employ suitable metrics for evaluating the utility of synthetic data generated by "Fusionstrap" and comparing them with those produced by the CTGAN and Synthpop methods.

RQ1.2: To understand to what extent the "Fusionstrap" method stands out compared to other data synthesis techniques, such as CTGAN and Synthpop, it is necessary to perform a comprehensive and comparative evaluation of the performances of these methods. The comparison will focus on the level of usefulness of the synthetic data generated by each method, with the aim of highlighting the possible advantages brought by "Fusionstrap".

RQ1.3: Managing class imbalance in datasets is a crucial problem in the field of machine learning. To determine whether "Fusionstrap" manages to make significant improvements in solving this aspect over the CTGAN and Synthpop methods, we will analyze in detail to what extent each method manages to balance the distribution of classes in the generated data sets. This comparison will provide essential insight into the effectiveness of the proposed approach.

RQ2: To what extent can "Fusionstrap" ensure the confidentiality of the original data?

RQ2.1 How to quantify the disclosure risk of synthetic data?

RQ2.2 How well does "Fusion strap" protect the confidentiality of the original data compared to other methods (CTGAN and SYNTHPOP)?

To answer this research question, we will perform specific analyzes to evaluate the privacy level provided by the "Fusionstrap" method and the other synthetic data generation methods (CTGAN and Synthpop). We will apply methods and metrics to quantify the disclosure risk of the synthetic data generated by each method.

RQ2.1: To quantify the risk of disclosure of synthetic data, we will use various measures and indicators that can be found in the specialized literature and evaluate to what extent the "Fusionstrap" method protects confidentiality.

RQ2.2: To evaluate how well “Fusionstrap” protects the privacy of the original data, we will perform a comparison with the other methods (CTGAN and Synthpop) based on the results obtained from the disclosure risk analysis. We will identify the possible advantages of the “Fusionstrap” method in ensuring confidentiality and evaluate whether it manages to provide a higher level of protection for the original data. To carry out these analyzes and comparisons, we will use an appropriate methodological framework and analyze the synthetic datasets generated by each method as well as the original data. This approach will allow us to objectively evaluate the ability of the “Fusionstrap” method to protect data privacy and provide a viable and more secure alternative in synthetic data generation.

Table 1 provides a concise summary of the research questions alongside the corresponding methods employed to attain the answers. Each research question is associated with the relevant research method, and the anticipated results of the research questions are also delineated.

Table 1

Research questions and methods

Sub-research Question	Research method	Outcome	Chapter
RQ1.1 How can the utility of synthetic data be measured?	Literature review	Definitions, evaluation method, evaluation metrics, evaluation tool	3;4
RQ1.2 What is the utility level of the “Fusionstrap” framework compared to other data synthesis methods (CTGAN and SHYNTPOP)?	Experiment Comparative analysis of the results	Comparative table of the evaluation results	5;6
RQ1.3 To what extent does “Fusionstrap” resolve class imbalances compared to CTGAN and SYNTHPOP?	Experiment Comparative analysis of the results	Comparative table of the evaluation results	5;6
RQ2.1 How to quantify the disclosure risk of synthetic data?	Literature review	Definitions, evaluation method, evaluation metrics, evaluation tool	3;4
RQ2.2 How well does "Fusionstrap" protect the confidentiality of the original data compared to other methods (CTGAN and SYNTHPOP)?	Experiment Comparative analysis of the results	Comparative table of the evaluation results	5;6

2.2 Research methods

As suggested by the standard approach to scientific research [13], this thesis follows four essential steps: preliminary research, conceptualization, hypothesis testing, and validation (Figure 1). In the first step, we aim to investigate the problem of class imbalances in datasets and focus on the solution of generating synthetic data to remedy this problem. The conceptualization stage involves gathering and synthesizing relevant knowledge from the literature, including techniques for generating synthetic data, methods for handling class imbalances [3], approaching the use of Bootstrap techniques [14] [15], and evaluating usability and privacy of the synthetic data [16]. Hypothesis testing consists of applying the “Fusionstrap” method to three distinct data sets: the US Census, diabetes prediction, and AIDS cases. In parallel, we will use the CTGAN and Synthpop methods to generate synthetic data from the same datasets. Evaluation of the quality of the synthetic data generated by each method will focus on utility and confidentiality. The usefulness of the synthetic data will be measured by analyzing appropriate metrics, in accordance with the specialized literature. In addition, we will assess the disclosure risk of synthetic data to quantify the level of privacy provided by each method. The final stage consists in the validation of the results, with the formulation of conclusions regarding the performance of the "Fusionstrap" method in the generation of synthetic data and in the management of class imbalances. For this purpose, we will compare the results obtained by the "Fusionstrap" method with those obtained by the CTGAN and Synthpop methods and evaluate whether "Fusionstrap" can provide a better and more realistic solution for synthetic data generation that combines data utility and privacy.

In the first and second stages, a thorough literature research will be conducted to understand the context of synthetic data generation as a solution for solving class imbalances in data and to develop a solid theoretical foundation. Stages three and four will consist of the practical part of the research, where experiments will be carried out to evaluate the performance of the "Fusionstrap" method to generate synthetic data and to solve class imbalances. Throughout this process, special attention will be paid to ensuring the validity of the results by identifying and addressing potential threats to it.

2.2.1 Literature review approach

In order to gain a comprehensive perspective on the field of synthetic data generation and class imbalance resolution, we will conduct a multivocal literary review (MLR). This approach entails broadening the scope of the systematic literature review (SLR) by encompassing gray literature in the search, in addition to the inclusion of published scientific literature [17]. The implementation of MLR will enable us to discern the existing knowledge on the subject and pinpoint areas that necessitate further investigation in our study.

To analyze the scientific literature, we will follow the following guidelines [18]:

- Search strategy: Our search strategy will involve using high-quality academic resources to gather information relevant to our research. We will access prestigious scientific databases, such as IEEE Xplore, PubMed, Google Scholar, Mendeley and ResearchGate, to identify scientific articles, journals, and conference proceedings relevant to generating synthetic data, handling class imbalances, and evaluating data utility and data privacy. We will also use the digital libraries of prestigious universities such as Harvard University, MIT (Massachusetts Institute of Technology) and Utrecht University to access relevant doctoral theses and research papers. Through this approach, we ensure that we have access to the most recent and authoritative information in our research field. Key search terms will be derived from the research questions and the specific context of our topic. We will use keywords such as "synthetic data generation", "class imbalance management", "bootstrap", "data utility evaluation methods" and "synthetic data privacy".
- Snowball method: We will also snowball forward and backward to identify additional relevant sources. This method involves identifying sources from the bibliography of relevant articles to obtain more detailed information and to discover other important works in our research field.

The review of existing literature will establish a robust basis for comprehending the present state of research in the domain of synthesizing data and handling class imbalance. This comprehensive examination

will enable the identification of current methods and techniques, along with their associated challenges and limitations. Additionally, it will pinpoint specific areas where our proposed method can effectively tackle class imbalance issues, generating synthetic data of superior quality and utility for machine learning models.

2.2.2 Experimental Research Methods

The experimental stage tests the hypothesis that “Fusionstrap”, the method proposed in this thesis, can be successfully used to generate synthetic data. This endeavor aims to provide a new perspective on synthetic data generation and demonstrate the utility and effectiveness of this approach in addressing class imbalances and maintaining data privacy.

To perform the experiment, we will follow the following steps:

- *Selection of data sets and research design*: We will choose three distinct data sets to evaluate the effectiveness of the Fusionstrap method in various contexts. This selection will be done carefully to ensure their representativeness and relevance in the fields studied. We will use a pre-experimental step to investigate how Fusionstrap can successfully address class imbalances in these datasets.
- *Synthetic data generation and real data analysis*: We will apply the “Fusionstrap” method on each data set to generate synthetic data. In parallel, we will perform analysis of real datasets to identify class imbalances and privacy vulnerabilities. This real data will serve as a benchmark for evaluating the usefulness and privacy of the synthetic data generated by "Fusionstrap".
- *Assessing the usefulness and privacy of synthetic data*: We will utilize different approaches to assess the effectiveness of synthetic data, drawing comparisons with both authentic data and data produced by alternative methods such as CTGAN and SYNTHPOP. Additionally, we will quantify the confidentiality level of the synthetic data and identify potential risks linked to information disclosure.
- *Comparison of results*: The results obtained by applying the "Fusionstrap" method will be compared with those obtained by using other synthetic data generation methods. We will analyze the performance of our method in resolving class imbalances and maintaining confidentiality to objectively evaluate the contribution of this research to the field of synthetic data generation. Also, to assess the usefulness of the synthetic data and to frame the experiment in a wider scientific context, we will compare the results obtained in the evaluation of two statistics from the AIDS data set with the results obtained by other researchers who have generated synthetic data using a method called "Avatar" on the same AIDS data set. This comparative analysis will enable a thorough evaluation of the "Fusionstrap" method, placing it in context alongside other established approaches to synthetic data generation.

Overall, this experimental process will provide robust data and relevant results to answer our research questions and make a significant contribution to the field of synthetic data generation with a focus on managing class imbalances and ensuring data privacy.

Summary:
In this research, we will combine a literature review with an extensive experiment to gain a comprehensive insight into addressing the class imbalance problem through the “Fusionstrap” method and to validate its contribution to the field of synthetic data generation.
The following chapter delves into the findings of the comprehensive review of scientific and related literature, which served as a crucial step in identifying and understanding the research problem at hand. This chapter aims to explore the existing body of knowledge and research in the field, providing valuable insights and context for the subsequent chapters.

2.3 Validity evaluation

In any scientific research, it is imperative to assess and address potential threats to the validity of the results. These threats can affect the quality, relevance and credibility of research, jeopardizing its validity. Within this context, we have identified and categorized potential challenges to the validity of our research thesis. This section provides a comprehensive analysis of these threats, classified into three main categories: internal threats, external threats, and conclusion threats, each with specific subcategories and outlined strategies employed for their treatment and management, thereby ensuring the validity of our research.

Internal validity:

1. *Data Quality*: The underlying data may contain errors or inaccuracies that distort the results. Using incorrect data may lead to wrong conclusions. To mitigate this problem, data cleaning has been performed to minimize errors and uncertainties. Cleaned dataset verification and validation were prioritized, by adding a statistical comparison after replacing missing values.
2. *Choice of Metrics and Parameters*: Subjectivity may arise in the choice of evaluation metrics and specific parameters for analysis. To mitigate this threat, we have selected objective metrics and documented decisions related to metrics to make them transparent.
3. *Methodological Limitations*: Specific limitations of the method used may influence validity. To address this issue, we have been transparent about the method and approach used, presenting the limitations openly.
4. *Threat of Subjectivity*: Subjective interpretation of results can be an internal threat. To address it, we used objective methods and techniques and ensured a rigorous discussion of results to avoid subjectivism.

External validity:

1. *Size of the Data Set*: The size of the data set used in the research may be insufficient to properly represent the entire population or to generalize the results to other data sets. Also, the relative size of the data set can influence the results. Using an insufficiently large or inappropriate data set can lead to inaccurate results or unrepresentative conclusions. To manage this threat, we selected three varied data sets, including adult census datasets, a diabetes prognostic dataset and an HIV/AIDS dataset, covering a wide range of sizes from 40,000 records to 2,000 records. This diversity of dataset sizes allowed us to evaluate and compare results across a diverse spectrum of contexts. Thus, we ensured that our research provides a comprehensive and generalizable picture of the issues addressed.
2. *Generalization of Results*: There is the threat of generalization of results to other domains or contexts, which is an external threat. To mitigate this problem, care was taken to explicitly mention the specific context of the study and possible limitations in generalizability.
3. *Impact of Algorithms*: The choice and configuration of algorithms can affect validity in an external context. To address this threat, transparency was ensured regarding the algorithms and parameters used, thus facilitating replication and validation of the study in other settings.

Conclusion validity:

1. *Reproducibility of the Study*: To ensure the validity of the conclusion, attention was paid to the reproducibility of the study. Full details of the methods and procedures used were documented to enable other researchers to replicate and validate the findings in different contexts.
2. *Consistency with Initial Objectives*: The validity of the conclusion was assessed against the original objectives of the research. Any significant deviation of the results was explained and justified in relative to the direction and purpose of the study.
3. *Relevance in the General Context*: To ensure the external validity of the conclusion, emphasis was placed on the relevance of the results in a wider context. The applicability and significance of the findings in similar or different contexts were considered.

4. *Consistency with Existing Literature*: The validity of the conclusion was strengthened by examining the coherence and consistency of the results with existing literature and research in the field.
5. *Adjustment for Limitations*: Limitations of the study were acknowledged and addressed in the context of the conclusions. Necessary adjustments or clarifications were made to account for the possible influences of these limitations on the validity of the conclusions.

These internal, external and conclusion threats were addressed by taking appropriate preventive measures and ensuring detailed documentation of the research process. Attention was also paid to the limitations of the study, thus helping to strengthen the validity of the results.

Chapter 3
Theoretical Background

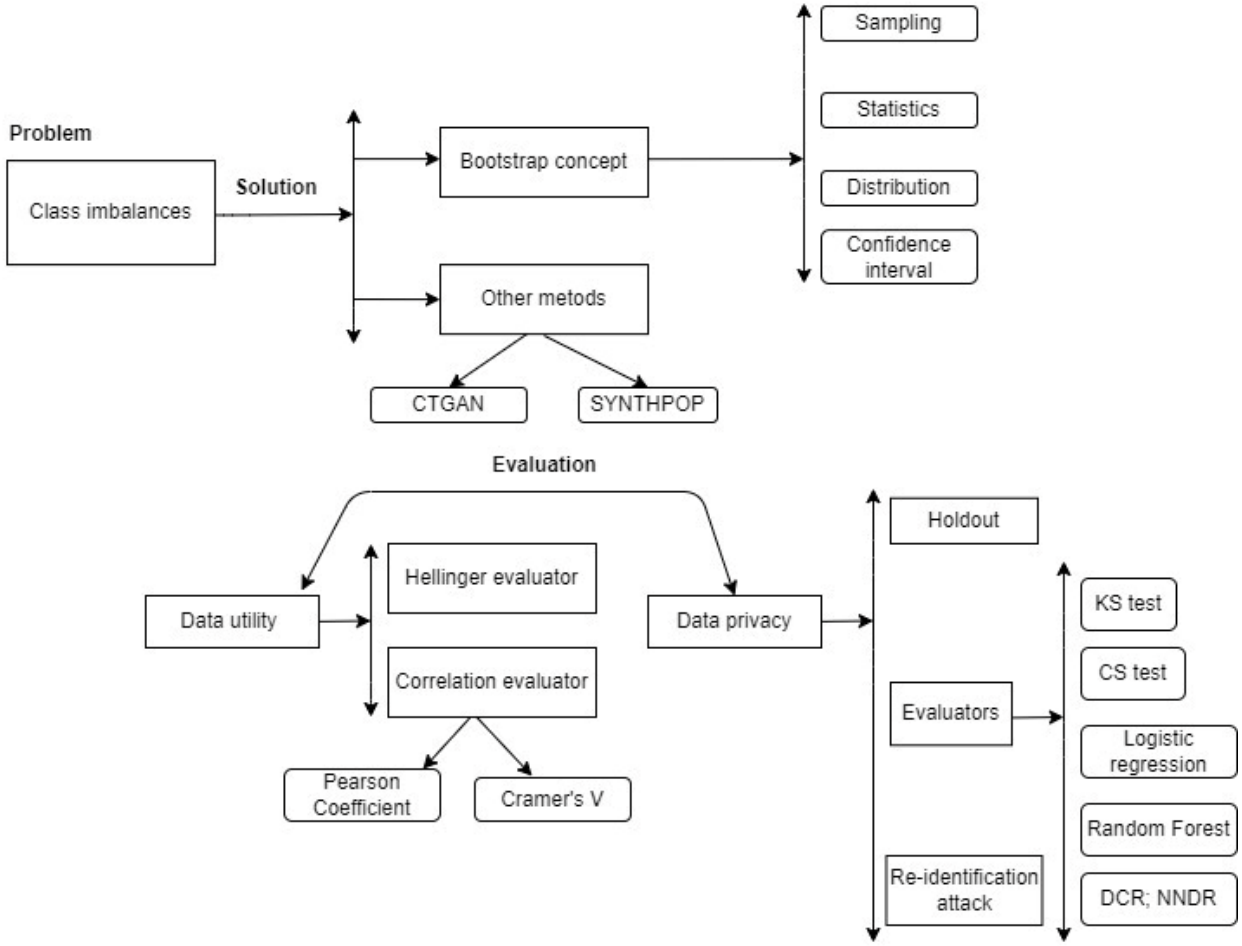
This chapter aims to analyze and present the essential conceptual framework for understanding our research. In this context, we will carefully review the literature relevant to our research area, exploring previous studies and research that address similar topics and that can provide context and support for our scientific objectives.
To identify the scientific pieces of information needed for the research, we mainly used Google Scholar, a search engine known for its ability to provide a diverse range of relevant articles. To minimize the potential limitations of relying on a single source, we also used other metadata services such as Mendeley. Mendeley brings into consideration readers' preferences, thereby providing a more comprehensive perspective on potential references for systematic literature reviews.
In order to precisely direct the source selection process in accordance with the objectives and the specific research area of this thesis, we applied the following criteria:

- We prioritized sources written in English, as this was essential to ensure the accessibility and quality of the research information.
- To ensure that we have the most recent information in the field, we focused especially on sources published after 2005. However, an exception was made for older sources that had relevance to our research field.
- We examined whether the sources provided information that directly answered one of our research questions. Each source was evaluated for its relevance to our proposed method, "Fusionstrap", and our research objectives. Sources that addressed or supported the key principles, techniques, or concepts underlying "Fusionstrap" were included. This was an essential step to build a solid theoretical basis for the development and evaluation of our method.
- Sources that directly correlate with our experiments have also been selected. This included sources that used the same data sets or addressed similar issues of class imbalance and generated synthetic data so that we could compare and contextualize our results.

In total, we analyzed 95 sources as part of our literary research. Of these, 20 sources focused on issues related to class imbalances, 5 sources addressed other methods of generating synthetic data, and 25 sources covered Bootstrap concepts. We also scrutinized 28 sources that discussed data utility issues and 17 sources that addressed data privacy issues. Of the 95 sources examined, 8 of them also included gray literature, which implies that we consulted various websites and platforms to gain a comprehensive perspective on the research topic.

Figure 2 schematically shows the structure of this section. It will explore issues related to class imbalances in datasets (Section 3.1), detail the concept of Bootstrap and its role in data analysis (Section 3.2), and provide a synthetic overview of two established methods for generating synthesis data, such as CTGAN and Synthpop (Section 3.3). In addition, concepts of data utility will be investigated (Section 3.4) and the issue of privacy of synthetic data will be examined in detail (Section 3.5). Through these analyzes and presentations, we aim to build a solid theoretical foundation for the development of our research, highlighting the key principles and concepts needed to explore the field and address our scientific goals.

Figure 2
Schematic theoretical background



3.1. Class imbalances

Class imbalances are a common problem encountered in datasets, where the distribution of classes is uneven, with some classes having significantly more examples than others. This discrepancy in the number of examples can significantly affect the performance of machine learning and data analysis models [19].

3.1.1 Causes and consequences of class imbalances

The causes of class imbalances can be varied and depend on the scope. Some of the common causes include:

- Rarity of events: In some domains rare events may be underrepresented in datasets. This can be the cause of significant class imbalances, as rare events are naturally less common. Maalouf Maher and Theodore B. Trafalis [20] emphasize the importance of sparse data in areas where such events can have a significant impact, such as fraud detection, medical diagnosis for rare diseases, or prediction of rare events such as major accidents. It also highlights the challenges of machine learning in the presence of sparse data, such as poor performance of models due to underrepresentation of minor classes, overlearning, and inaccurate classification results. In addition, it emphasizes the need to develop appropriate methods and approaches to obtain accurate and relevant results in the case of these unbalanced datasets and presents various approaches and techniques used to deal with sparse data, including subsampling, instance overlap, and generation of synthetic data.

- Biased Sampling: Data collection methods may be subject to choices that lead to incorrect representation of classes. Sampling bias occurs when the sampling of data is not representative of the entire population, which can distort the results of the analysis. For example, in medical studies, patients with severe symptoms may be more likely to be included in the data set, resulting in a majority class of patients with severe symptoms and a minority class of patients with milder symptoms. Sampling error and class imbalance have the potential to impact the performance of logistic regression models, resulting in imprecise estimates and inaccurate, uninterpretable results [21]. To address these issues and obtain accurate and robust estimates of logistic model parameters, weighted sampling methods and synthetic data generation techniques can be used [21].
- Labeling errors: In large and complex data sets, there is a possibility of labeling errors or confusions, which can affect the correct proportion of classes in the data set. These errors or confusions can occur for a number of reasons, such as human error in the labeling process, inaccurate automatic labeling, or ambiguities in the definition of classes. The impact of labeling errors can be significant and lead to distortion of the correct distribution of classes in the data set. This can negatively affect the performance of machine learning models leading to incorrect and unreliable results, as well as a lack of generalization in the classification and prediction of new data, as models can be trained to learn from errors based on wrong labels [22]. Therefore, identifying and solving labeling errors in the data preprocessing stage can have a significant impact on the quality and generalization of machine learning models in the face of new data.

3.1.2 Approaches and methods for solving class imbalances

Addressing and resolving class imbalances are crucial aspects in handling data sets characterized by unequal class distributions. There are multiple techniques and methods used to address this problem and ensure a fair and accurate classification or prediction. One of the most common approaches to resolving class imbalances is dataset rebalancing [25]. This involves adjusting the distribution of classes in the data set to achieve greater balance. There are four main methods of rebalancing:

3.1.2.1. Oversampling

Oversampling is a method used to deal with class imbalances by adding new instances of minor classes until the number of instances in each class becomes balanced. This approach has the advantage of keeping all the data in the minority class, but may increase the risk of overfitting and learning noise in the dataset.
SMOTE (Synthetic Minority Over-sampling Technique) [23] is a popular oversampling technique that generates new synthetic examples for the minority class by interpolating between existing neighbors. The process involves randomly selecting an example from the minority class and identifying its k nearest neighbors. Then, two points are randomly chosen from this group and their weighted average is calculated. This weighted average represents a synthetic sample that is added to the data set.
ADASYN (Adaptive Synthetic Sampling) [24] is a variant of the SMOTE method that adjusts the degree of oversampling according to the classification difficulty of each example in the minority class. Hence, ADASYN produces a greater number of synthetic examples for instances in the minority class that pose a more challenging classification task, while generating fewer examples for those that are easier to classify.
Random Oversampling is a simple oversampling technique that consists of randomly copying samples from the minority class until balance between classes is reached. This method can be easily applied, but can lead to overfitting and increased variance in the data set [25].
These oversampling techniques can be used individually or in combination to obtain a balanced data set.

3.1.2.2 Under-sampling

Under-sampling is an approach to dealing with class imbalances by randomly or strategically removing examples from the majority class so as to obtain a data set with more balanced proportions between classes. Among the under-sampling methods are:

- Random Under-sampling: This technique entails randomly eliminating some instances from the majority class to address class imbalances. Nonetheless, a drawback of this approach is the potential loss of crucial information from the majority class, which can impact the model's performance [29].
- Tomek Links: This method involves identifying the pairs of examples from the majority and minority classes that are closest to each other. Then, examples from the majority class that pair with examples from the minority class are removed. This removes only those examples that are in the conflict zone between the two classes, which can lead to better separation of classes and improved model performance [30].
- Edited Nearest Neighbors (ENN): This technique entails examining the nearest neighbors for each sample within the dataset. Instances in the majority class that predominantly have neighbors from the majority class will be eliminated. In essence, ENN removes those instances from the majority class that are deemed to be in close proximity to the minority class. This process aids in balancing the class proportions and enhancing the performance of the model [31].

3.1.2.3 Methods based on cost-sensitive learning

Methods based on cost-sensitive learning are approaches that focus on handling class imbalance by assigning different costs to major and minor class classification errors. These methods put more emphasis on correctly classifying minor classes, and their classification errors are penalized more than those of major classes. Their goal is to encourage machine learning models to pay more attention to smaller classes to achieve more balanced and accurate results. There are several techniques and algorithms that can be used to implement cost-sensitive learning. Some of these include:

- Cost-sensitive decision trees: These are variants of decision trees that take into account the different costs for correctly and incorrectly classifying classes. Examples of such methods include Cost-Sensitive C4.5 and Cost-Sensitive Random Forests [26].
- Cost-sensitive support vector machines (SVM): In this case, the cost function is optimized to obtain a decision boundary that minimizes cost-sensitive classification errors. Some examples of such methods are Weighted SVM and SVM with asymmetric penalty [27].
- Cost-sensitive logistic regression: This is a variant of logistic regression that uses cost matrices to modify the cost function and treat the correct classification of minor classes more carefully [28].

3.1.2.4 Stratified Bootstrap

The stratified bootstrap is a method used to solve class imbalances, being applied both as an oversampling method, by stratified resampling in each class to obtain a synthetic data set balanced in terms of class distribution, and as an under-sampling method, by strategic or random removal of examples from the majority class to achieve class balance [32], [33].
Pertami J. Kunz and Abdelhak M. Zoubir explore the use of the stratified Bootstrap method in the context of training a tampered food detector [34]. In this paper, the authors focus on the detection and resolution of class imbalances in the dataset, so that the detector is able to accurately recognize adulterated foods and minimize the risk of classification errors. The dataset used in the paper shows class imbalances, where classes representing adulterated foods are underrepresented compared to classes representing unadulterated foods. To overcome these imbalances, the authors propose the use of the stratified Bootstrap method, which allows resampling based on data heterogeneity layers [34]. This approach ensures that the proportions between classes are balanced in bootstrap-generated datasets. The experimental findings demonstrate a notable enhancement in the performance of the adulterated food detector due to the implementation of the Stratified Bootstrap method. By using this method, a more accurate classification of adulterated foods was obtained, minimizing the risk of errors and false positives [34].
“Fusionstrap” generates synthetic data by approaching a hybrid algorithm based on Stratified Bootstrap and Gaussian Copula Synthesizer. Chapter 4 includes a detailed description of this method.

3.2. Bootstrap concept

In this section, we will investigate in detail the fundamental concept of Bootstrap, an essential method in data analysis, how it is applied in research, and examine both its benefits and limitations.

3.2.1 Definition and rationale for using Bootstrap in data analysis

Bootstrap is a statistical resampling method used to estimate probability distributions and statistics of a data set. It was introduced by Efron in 1979 as a technique for obtaining robust estimates in data analysis [35]. The main idea behind Bootstrap is to estimate the sampling distribution by repeating sampling with replacement from the original data set, thereby obtaining multiple bootstrapped samples [36]. These bootstrapped samples are then used to compute the statistics of interest and to estimate the confidence interval of these statistics [37].
The Bootstrap method offers a straightforward and potent technique for estimating probability distributions and statistics, alleviating the necessity to make assumptions about the underlying data distribution [36]. Its utility becomes particularly apparent when the distribution of the data is either unknown or defies description through conventional statistical methods [38]. Through iterative sampling with replacement, Bootstrap effectively captures the inherent variability in the data, resulting in robust estimations of the outcomes [37].

3.2.2. Bootstrap's base method

The basic Bootstrap method, originally proposed by Bradley Efron in 1979, is a statistical technique used to estimate the probability distribution of a sampling statistic [35]. It has been and is frequently used in data analysis and statistical inference [36]. The Bootstrap algorithm highlights two fundamental components: Resampling and Bootstrap statistical estimation (Figure 3).

3.2.2.1 Resampling procedure

The Bootstrap resampling procedure has the following key features:

- Sampling with replacement: This method involves resampling a single original data set, whether it represents a population **(Step 1 in Figure 3)** or a representative sample from that population **(Step 2 in Figure 3)** thus obtaining several bootstrap samples **(Step 3 in Figure 3)**. When drawing elements from the original sample to form the re-sampling samples, each element is replaced in the original sample before drawing. The bootstrap method assigns an equal probability to the random selection of each data point from the original sample, ensuring their inclusion in the resampled datasets. As a result, some records will be sampled multiple times in bootstrap samples, while others will not be sampled at all. This property defines the "with replacement" expression of the process [39].
- Bootstrap sample size: The process generates resampled datasets matching the size of the original dataset, comprising diverse combinations of values. This occurs as each sample involves drawing with replacement from the original sample. Each simulated dataset possesses a distinct set of descriptive statistics, including mean, median, variance, and standard deviation [40], [41].
- Number of Bootstrap Samples: The number of bootstrap samples is a crucial component in the resampling process and is chosen to ensure an accurate estimate of the probability distribution. The theoretical minimum number of bootstrap samples can vary and is generally determined by the need to cover a representative range of the underlying distribution. Although Efron and Tibshirani [40] mention 25 samples as a possible value, in practice this theoretical minimum can be influenced by the size of the data set and the degree of variability in the data. The theoretical maximum number of bootstrap samples is a more flexible concept and may depend on available resources, computing time, and desired accuracy. Usually, in the literature [42], it is recommended to use several hundred or even thousands of samples to obtain results close to the ideal bootstrap. However, there is always a trade-off between a theoretical maximum number and a practical one, since a larger number of samples implies an increased computational cost.

3.2.2.2 Estimation of distributions and statistics by Bootstrap

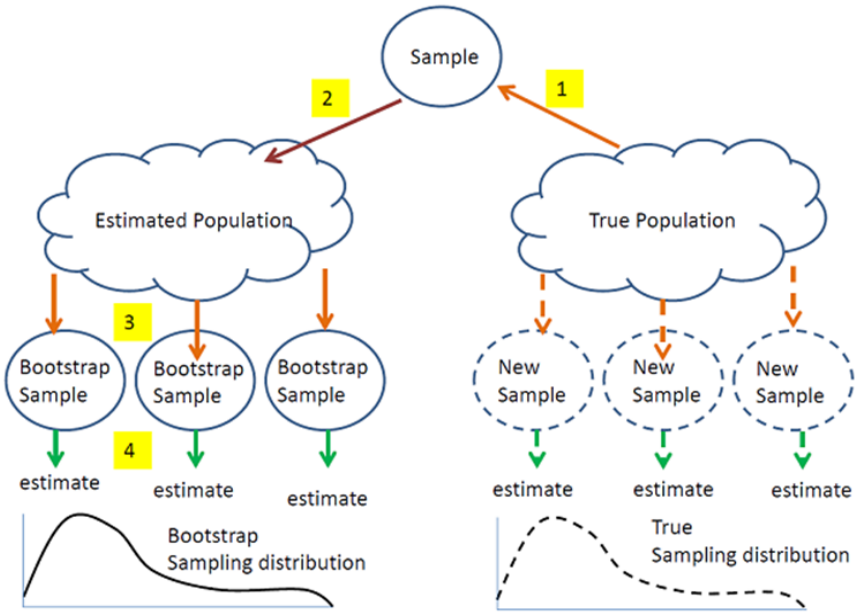
The basic Bootstrap technique has become a popular method of statistical inference due to its ability to estimate probability distributions and allow confidence intervals to be obtained without the need for assumptions about the distribution of the data [42], [43].

Bootstrap statistical estimation is a nonparametric method of estimating the distribution of a statistic of interest using repeated resampling of samples from a data set. This procedure entails creating an extensive set of bootstrap samples through the random selection of data with replacement from the original dataset, followed by the computation of the desired statistics for each sample. Resampling in this manner proves valuable when the precise shape of the population distribution is uncertain and when the sample size is constrained. It is a robust and non-parametric technique that can be applied to estimate parameters and statistics in various research fields [40], [42]. Most commonly, these include the standard error and variance of a population parameter (e.g., a mean, median, correlation coefficient, or regression coefficient) [53].

Distribution estimation entails assessing data acquired through resampling to create approximations of the distribution. Utilizing the samples obtained through resampling, calculations for estimates of the statistic of interest (such as mean, median, standard deviation) can be performed (Step 4 in Figure 3) and confidence intervals for these estimates can be obtained [41], [42]. These statistics provide estimates of the distribution. To assess the variability of these estimates, a distribution of the respective statistics is constructed, providing insight into how these estimates may vary across different samples. It is essential to highlight that this procedure constitutes a core element of the Bootstrap method, grounded in the concept that estimates derived from the resulting samples frequently exhibit a Gaussian distribution [45]. Bootstrap essentially treats the sample as if it were the entire population.

Figure 3

The steps of the nonparametric bootstrapping process



Note. From [44]

For the correct application and interpretation of the Bootstrap method and, finally, for obtaining valid conclusions in the analysis of the results, we will further define the concepts of population, sample, parameters, statistics and statistical inference.

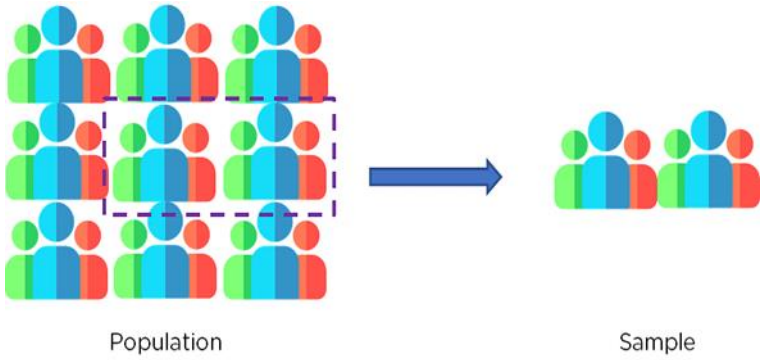
Definition 1 (Population). In statistical terms, a population refers to the total number of statistical units (individuals, objects, events) that share at least one common characteristic and are the focus of interest for a statistical analysis [46]. The population can refer to a set of units that either currently exists or is a

conceptual group. For example, a sample population might include all students attending a school at the time of data collection. The data collection process involves gathering information from each of these students, depending on the research objectives (Figure 4).

Definition 2 (Sample). The sample constitutes an ostensibly representative subset of a population chosen through a specified procedure [46]. For example, a sample may be a random subset of 20 students selected from the population for data collection (Figure 4). In statistical testing, a sample is employed when the size of the population is impractical for all members or observations to be included in the test. The outcomes derived from a representative sample participating in a study can be extrapolated to make generalizations about the entire population.

Figure 4

Sample from a population



Definition 3 (Parameter vs. Statistics). A parameter is a numerical representation characterizing an entire population, such as the population average. Conversely, a statistic is a numerical representation characterizing a sample, like the sample average [52]. Examples of common parameters of the population of interest and the corresponding sample statistics can be:

Table 2

Parameter vs. Statistics

Quantity	Parameter	Statistics
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s
Proportion	p	p^

One of the "statistics" of the sampler can also be called "an estimator". As an illustration, the sample mean $\bar{x}$ serves as an approximation of the population average. An estimator (T) is a function of random variables, and therefore, it is itself a random variable, which provides a way to estimate T for the entire population. A star next to a statistic, such as T^* (e.g., s* or $\bar{x}$*), indicates that the statistic was calculated by re-sampling.

Definition 4 (statistics). A statistic is a function of observable random variables, determined by a probability distribution without any unspecified parameters, as indicated in reference [47].

Definition 5 (statistical inference). Statistical inference is defined as the procedure for analyzing the result and making decisions about the parameters of a population resulting from random sampling. Statistical

inference aims to gauge the uncertainty or variability across different samples [47]. This process commonly leans on the sampling distribution and the standard error of the characteristic under consideration.

Definition 6 (Variance and Bias): Variance is the measure of dispersion or difference between individual values and the mean of a data set, and low variance estimators are preferred because they have greater stability and provide more accurate estimates of population parameters. Concurrently, bias denotes the systematic disparity between the estimated mean value and the true value of the population parameter, potentially manifesting in certain estimates [56].
Bootstrap methods can help us estimate the variance and quantify the bias associated with a sample-based statistic [54]. By re-sampling with replacement from the original data set, Bootstrap creates replicas of the samples, which allows us to obtain an approximation of the distribution of the respective statistic and assess its variability [56]. This allows us to obtain an approximation of the distribution of the statistic and calculate the standard errors associated with our estimates [55]. We can also assess the variability of the results and construct confidence intervals to indicate the level of uncertainty of our estimates [51]. To form a confidence interval, it is necessary to measure the variability of the initial sample statistic. For instance, when determining a confidence interval for the population mean, it is crucial to estimate the expected variation of the sample mean across different samples. Bootstrap can provide an estimate of this variability by generating multiple samples and evaluating the dispersion of the statistic based on them [54].

For a data set with n values (x_1, x_2, ..., x_n), the variant is calculated in several steps:

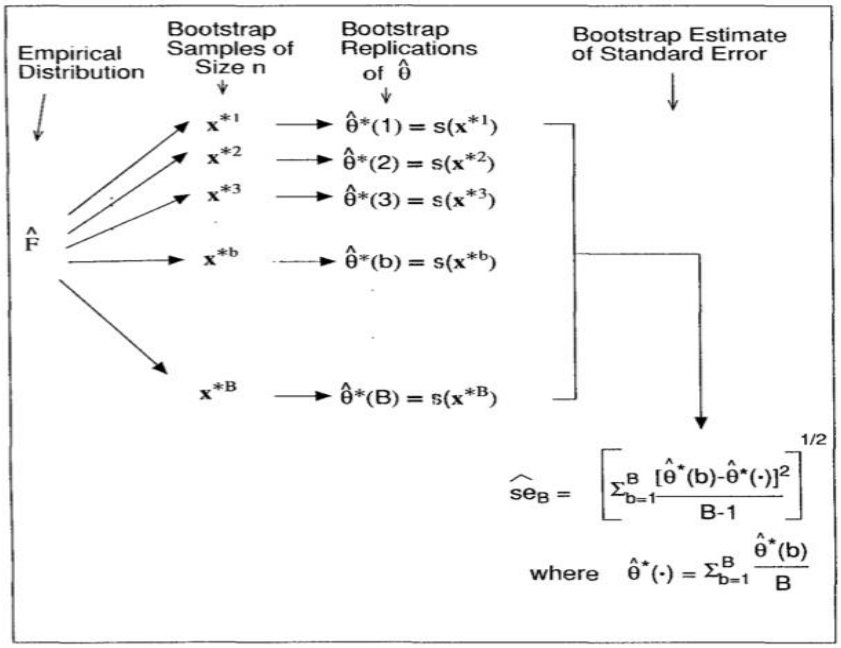
1. The average ($\bar{x}$) of the data set is calculated: $\bar{x} = (x_1 + x_2 + ... + x_n) / n$
2. Calculate the difference between each value and the mean: $(x_1 - \bar{x})$, $(x_2 - \bar{x})$, ..., $(x_n - \bar{x})$
3. Square each difference: $(x_1 - \bar{x})^2$, $(x_2 - \bar{x})^2$, ..., $(x_n - \bar{x})^2$
4. The average of the squares of the differences is calculated: $Var = [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2] / n$

Variance is a metric that quantifies the spread or deviation of data points within a dataset in relation to its mean. The larger the variance, the more dispersed the individual values in the data set are and deviate significantly from the set mean. This dispersion indicates increased heterogeneity in the data, suggesting that it may cover a wide range of values. Conversely, with a smaller variance, individual values tend to be closer to the mean of the dataset. In this scenario, the data is perceived as more homogeneous, signifying that the majority or all of the values in the set cluster around the mean. A lower variance level may indicate a more uniform distribution in the data.

Definition 7 (standard error). The standard error (SE) quantifies the spread or deviation of a statistical estimate from the mean value of that estimate [36]. It indicates the precision of a statistical sample and mirrors the variability among the estimates we would derive if we repeatedly sampled the population. In essence, it quantifies how much our estimate deviates from the true mean value of the characteristic or parameter we aim to estimate.
Although conventional approaches are deemed sufficient for computing the standard error of a sample statistic, the bootstrapping method employs a replacement technique, generating multiple standard error values that collectively represent the mean SE [40]. As an illustration, when estimating the standard error of the mean using Bootstrap (Figure 5), the process involves computing the mean for each bootstrap sample, determining the mean of the initial sample, and subsequently calculating the variance of the means obtained in these steps to estimate the variance of the sample mean [40]. The standard error of the sample mean is then derived as the square root of the variance obtained in the preceding step. This methodology provides a means to gauge the accuracy of the sample mean as an approximation of the population mean. A smaller standard error indicates a more precise estimate of the population mean.

Figure 5

Algorithm for estimated standard error of a statistic



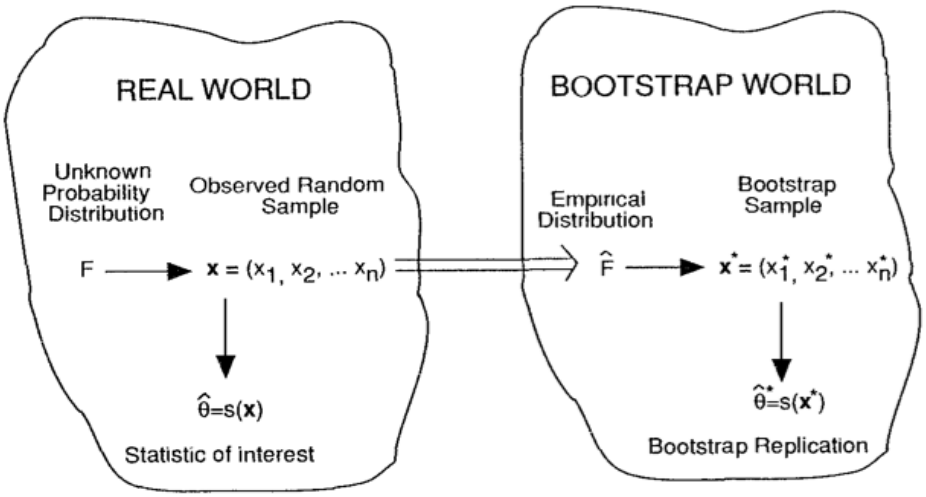
Note. From [40, p. 48]

The bootstrap simulation error, which measures the disparity between the actual distribution and the estimated distribution, encompasses two distinct errors originating from separate sources: a bootstrap (statistical) error and a simulation (Monte Carlo) error, as expounded by Zoubir A. M. and Iskandler D. R. [54]. The authors contend that the first error is inevitable and is independent of the number of Bootstrap samples (B), but it may be contingent on the size (n) of the original sample. On the other hand, the second error can be mitigated by augmenting the number of bootstrap samples. Consequently, the objective is to select an appropriate value for B, ensuring that the simulation (Monte Carlo) error does not surpass the bootstrap error. Nonetheless, as the original data size (n) increases, the bootstrap error tends to decrease. Scientific studies have determined that the general guideline of selecting B = 40n, as suggested by Davison and Hinkley [36], is suitable in numerous contexts. Additionally, the jackknife-after-bootstrap method [55] offers a means to evaluate the impact of each error, such as bootstrap error versus Monte Carlo error. In real-world applications, the choice of B depends on the specific context and is left to the experimenter's judgment.

Definition 8 (distribution function). The distribution function (F) describes how the units are distributed in a population/sample, i.e., it indicates the probability (frequency) with which a random variable takes a certain value [51]. Through an examination of the statistics derived from the Bootstrap samples, we estimate the probability distribution of the statistic of interest (Figure 6). This distribution captures the uncertainty inherent in estimating the statistic based on the original sample.

Figure 6

Diagram of the bootstrap as it applies to one-sample problems

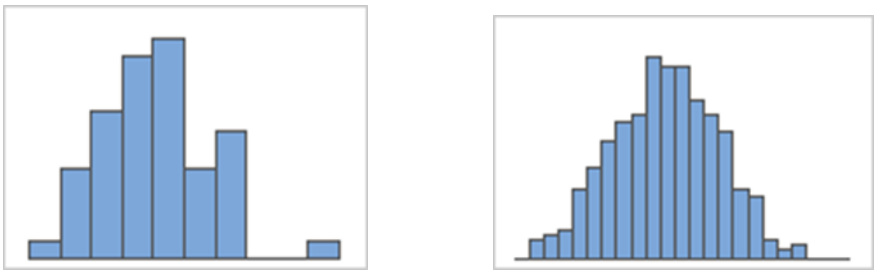


Note. From [40, p. 87]

Typically, a random variable is denoted by capital letters, such as X, representing a potential value that has not been realized. The likelihood of attaining a specific value is conveyed through its probability distribution. Once a value is observed, a lowercase letter, like x, is employed to differentiate it from the yet-to-be-realized random variable X. The observed value is not subject to randomness but signifies an actualization of a random variable. Thus, a sample of data drawn from a distribution F will be denoted x1, x2, . . ., xn (Figure 7). Likewise, a bootstrap sample is denoted with the "star" notation: x*1, x*2, . . ., x*n. This notation closely resembles the convention for representing sample data, typically expressed as: x1, x2, . . ., xn (Figure 7).

The sampling distribution represents a theoretical compilation of all potential estimates that would emerge if the population were subjected to repeated resampling (Figure 5). The underlying theory stipulates that, given certain conditions such as ample sample sizes, the sampling distribution will approximate normality, with the distribution's standard deviation equaling the standard error [48]. Nonetheless, situations arise where the sample size is insufficient, or the assumption of a normally distributed theoretical sampling distribution is untenable. This complexity hinders the determination of the standard error of the estimate and makes drawing meaningful conclusions from the data challenging. Bootstrap has emerged as a valuable tool in such scenarios, proving effective in identifying and visualizing the sampling distribution of a statistic (e.g., mean) or the parameters of a model (e.g., β1 or AIC in linear regression) [48]. With Bootstrap, the sampling distribution is constructed by repeatedly resampling observations N times, each resampled set comprising n observations with replacement, rather than relying on theoretical calculations. The sampling distribution can simply be observed, and no hypothesis needs to be formulated in advance. The benefit of employing multiple re-sampling lies in achieving a more accurate estimate of the sampling distribution. The bootstrap distribution, represented in Figure 6, encompasses the distribution of the selected statistic derived from each resampling iteration. Ideally, the bootstrap distribution should exhibit a normal appearance. If, however, the bootstrap distribution deviates from normality, it raises concerns about the reliability of the bootstrap results. The clarity of the distribution typically becomes more evident after numerous resampling instances. For instance, as illustrated in Figure 7, the distribution appears indistinct with 50 resamples, but becomes more consistently normal with 1000 resamples.

Figure 7
Bootstrap distribution



50 resamples 1000 resamples

Definition 9 (Confidence interval). Confidence interval estimation is an important component in statistical data analysis and plays a crucial role in the interpretation of results. This range of values is an interval that encompasses the true value of a statistic of interest with a certain level of confidence, indicating the probability of coverage [49]. Typically, the confidence interval is chosen to have a coverage probability of 95% or 99%. The resulting confidence interval is presented along with the point estimate of the statistic of interest. This range provides a measure of the uncertainty associated with the estimate and helps to interpret the results in a more robust way.
Bootstrapping can be used to obtain approximate confidence intervals for certain statistics of interest [50]. The bootstrap confidence interval is constructed using the percentiles of the bootstrap distribution to establish a range for the parameter of interest [51]. A common method for obtaining an approximately 100(1-α) percent confidence interval through bootstrapping is the Reflection method, also known as the Percentile method [42]. This approach involves extracting the lower 100 percentile (α/2) and the upper 100(1-α) percentile from the Bootstrap distribution βl^. For instance, to compute a 90% confidence interval, one needs to identify the 5th and 95th percentiles, encompassing 90% of the data in between. These two percentiles would be the endpoints of our confidence interval.
In some situations, the sampling distribution resulting from the bootstrap method often does not appear to be normal. This is because in nonparametric bootstrapping only certain numbers can be chosen from the distribution - those from the original sample. This results in large gaps in the sampling distribution. If one were to create a confidence interval based on a normal distribution using this information, the assumption of normality would be violated, and the confidence interval would not be correct. Efron [44] performed an adjustment of the confidence intervals created by bootstrap methods by introducing bias-corrected confidence intervals that accounted for the non-normality observed in the estimate of the bootstrap sampling distribution. He improved these intervals again in 1987 to create confidence intervals for BCa (also known as corrected and adjusted confidence intervals). These adjustments to the original method were successful and are still the main methods used today.

3.2.3 Advantages and Disadvantages of the Bootstrap

The Bootstrap method is a powerful and versatile technique used in statistical and inferential analysis. It offers numerous advantages as well as some disadvantages that must be considered in its application.
The **advantages** of the Bootstrap method include:

- *Non-parametric flexibility:* The Bootstrap method is non-parametric, indicating that it does not necessitate specific assumptions about the precise shape of the data distribution. This provides flexibility in applying the method to various statistical contexts, being suitable for data with unknown or complex distributions. The absence of specific requirements about distribution parameters makes Bootstrap adaptable to various types of samples, providing a robust approach.
- *Standard Error Estimates and Confidence Intervals*: The Bootstrap method enables the computation of standard errors and the construction of confidence intervals for diverse statistics derived from the samples. This capability provides essential information for assessing the precision of estimates, allowing researchers to quantify the uncertainty associated with the results obtained.
- *Robustness:* Bootstrap is recognized for its robustness to outliers in the data. By repeatedly resampling the data, the Bootstrap method provides more stable and robust estimates than traditional methods, meaning that the results are not heavily influenced by the presence of unusual or extreme observations.

Disadvantages of the Bootstrap method include:

- *Computational cost*: For large or complex data, the resampling process can become computationally expensive, requiring significant resources to achieve accurate results.
- *Underestimation of variance*: In some situations, Bootstrap can underestimate the variance of statistics, especially with highly skewed or long-tailed data.
- *Dependence on original data:* Bootstrap results can be influenced by the original sample, which can lead to misinterpretations if the sample is not representative of the population of interest.
- *Limitations for small sample sizes*: For very small samples, the Bootstrap method may provide inaccurate or unrepresentative estimates for statistics, limiting its usefulness in certain situations.

In conclusion, the Bootstrap method is a valuable and efficient technique to estimate standard errors, construct confidence intervals, and perform statistical inference without requiring strict assumptions about the data distribution. However, we must consider both its advantages and disadvantages in choosing and interpreting the results.

3.3 Methods of generating synthetic data

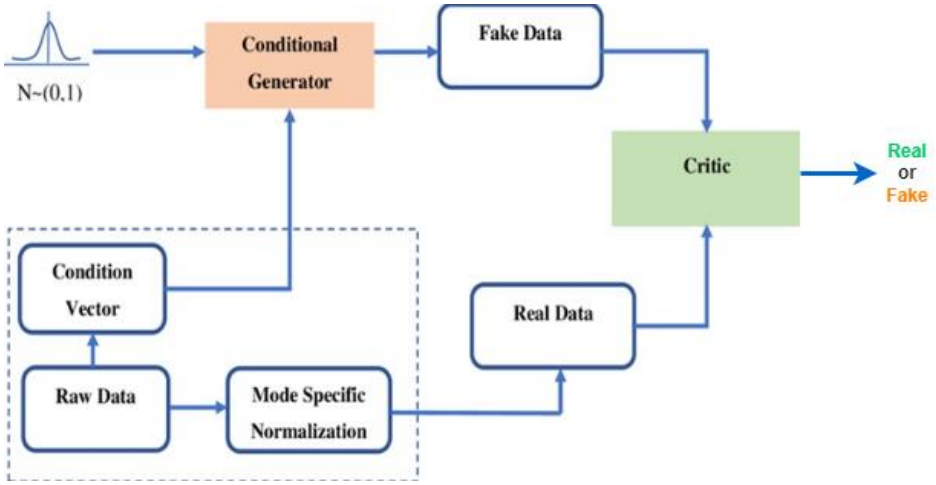
Within the context of addressing class imbalances, a potentially effective strategy involves the creation of synthetic data. This approach aims to rectify the class distribution imbalance and enhance the efficacy of machine learning algorithms. Two advanced methods for generating this synthetic data are CTGAN (Conditional Tabular Generative Adversarial Network) and Synthpop.

3.3.1 Conditional Tabular Generative Adversarial Network (CTGAN)

CTGAN leverages the Generative Adversarial Network (GAN) concept, employing a deep neural network architecture comprising two primary components (Figure 8): the generator and the discriminator [57], [58]. The generator's objective is to produce synthetic data that closely mimics the patterns observed in the real dataset. It achieves this by utilizing a series of random noise vectors and attempting to grasp the probability distribution inherent in authentic data, thereby generating synthetic data that exhibits a high degree of resemblance to reality [58]. The discriminator serves as a binary classifier, tasked with discerning between authentic and synthetic data generated by the generator. Its primary objective is to become adept at accurately classifying data and distinguishing synthetic data from genuine instances [58]. During the training process of CTGAN (depicted in Figure 8), a confrontation unfolds between the generator and the discriminator. The generator's objective is to deceive the discriminator by generating synthetic data that closely resembles real data. At the same time, the discriminator tries to get better and better at identifying synthetic data [58]. An essential characteristic of CTGAN is its ability to be conditioned on specific features or classes within the dataset. This implies the capacity to designate particular classes of interest and generate synthetic data in accordance with these specified classes [58].

Figure 8

CTGAN architecture



Note. Adapted from [58, p. 20533]

The advantages of the CTGAN (Conditional Tabular Generative Adversarial Network) method include:

- CTGAN can generate synthetic data that is highly realistic and resembles real data from the original dataset. This improves the quality and credibility of synthetic data.
- CTGAN can be trained to generate synthetic data taking into account the distribution and relationships between variables in the original data set. Thus, it can be ensured that the synthetic data preserves the relevant features and correlations.

CTGAN proves effective in addressing class imbalance by generating synthetic data for minority classes, thereby enhancing classifier performance.

Disadvantages of the CTGAN method include:

- CTGAN may require a relatively large dataset to operate efficiently and generate relevant and representative synthetic data.
- The training procedure for the CTGAN model can be time-consuming and computationally demanding, particularly when dealing with large datasets.
- As with any generative model, CTGAN can be susceptible to overfitting, which can lead to the generation of synthetic data with too much variety that does not adequately reflect the distribution of the real data.

3.3.2 Synthpop

Synthpop falls under the category of synthetic data generation methods based on Monte Carlo Markov Chain (MCMC) sampling. This method uses a sampling approach to generate synthetic data that respects the distribution and characteristics of the original dataset, and the generation process is driven by user-defined rules and constraints [59], [60].

The synthetic data generation process with Synthpop is customizable and takes place in two main steps: data preparation and actual synthetic data generation [59], [60].

Data preparation:

- In this step, the relevant variables from the original dataset are selected to be used to generate the synthetic data.
- Variables can be numeric or categorical, and Synthpop can handle both types of variables efficiently.
- Associations and correlations among variables are identified to ensure that the intricate structure of the original data is retained in the synthetic dataset.

Generating synthetic data:

- Synthpop uses a Monte Carlo Markov Chain (MCMC) sampling approach to generate synthetic data.
- The data generation process is driven by a series of user-defined rules and constraints that control the distribution of the synthetic data and preserve important features of the original dataset.
- For every data point in the original dataset, a synthetic data sample is created using the distributions of the pertinent variables and the defined constraints.

This tailored approach enables the generation of synthetic data that is realistic and relevant for further analysis.

Advantages of the Synthpop method include generating synthetic data tailored specifically to the original dataset, capturing complex relationships between variables, and control over the distribution of the synthetic data. This method is useful in cases where data sets are complex and have important features for further analysis.
Disadvantages of the Synthpop method include longer processing time for large data sets and potential overflow, which can affect the quality of the generated synthetic data.

3.4 Evaluation of synthetic data utility

The efficacy of synthetic data pertains to the proximity of synthetically generated data to the authentic data within a dataset. In the realm of synthetic data generation, the primary goal is to guarantee that the synthetic data retains the attributes and structure inherent in the original dataset. In other words, synthetic data must be representative and provide relevant information to support statistical analysis, model development, or other decision-making processes.
Assessing the usefulness of synthetic data is essential to ensure that these data are sufficiently representative and accurate to be used in analysis or decision making. When assessing the efficacy of synthetic data, various crucial methods and metrics enable the measurement of the compatibility between the synthetic dataset and the actual data. Examples include the Hellinger evaluator and correlation evaluators.

3.4.1 Hellinger Evaluator

In evaluating the utility of synthetic data, a key aspect is comparing the distributions of the synthetic data with those of the real data. Because there can be many variables in a data set, it is difficult to visually compare the distributions for each of them. To solve this problem, we can use summary statistical measures such as the Hellinger distance.
The Hellinger distance serves as a similarity metric employed for comparing two probability distributions. It revolves around the concept of computing a distance between the square roots of the probabilities within the two distributions, offering a quantification of the degree of similarity between them [61]. When considering two discrete probability distributions $P = (p_1, \ldots, p_k)$ and $Q = (q_1, \ldots, q_k)$, the Hellinger distance $H(P, Q)$ is defined as:

$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2}$$

where:

- H(p,q) represents the Hellinger distance between the distributions
- *pi* and *qi* are the probabilities associated with the values *i* from the two distributions.
- the sum is calculated for all possible values *i* from the two distributions

The major advantage of this metric is that it is limited and easy to interpret. The Hellinger distance is a probabilistic metric that spans from 0 to 1. A value approaching 0 implies high similarity between the

synthetic data and the real data, whereas a value nearing 1 signifies substantial disparities between the two distributions [61]. Regarding the generation of synthetic data, the Hellinger distance can be employed to evaluate the alignment between the distribution of synthetic data and that of real data. This aids in gauging the utility of synthetic data in comparison to the original dataset.

3.4.2 Correlation Evaluators

In assessing the efficacy of synthetic data, it is crucial to employ robust and informative metrics that gauge the congruence between the distribution of synthetic data and that of the real data. In this regard, metrics such as Pearson's coefficient, correlation ratio, and Cramer's V index are meaningful tools to assess the degree of similarity between synthetic and authentic data. These metrics provide clear and quantifiable insight into the effectiveness of synthetic data generation and can support informed decision-making regarding its use in further analysis.

3.4.2.1 Pearson Correlation Coefficient

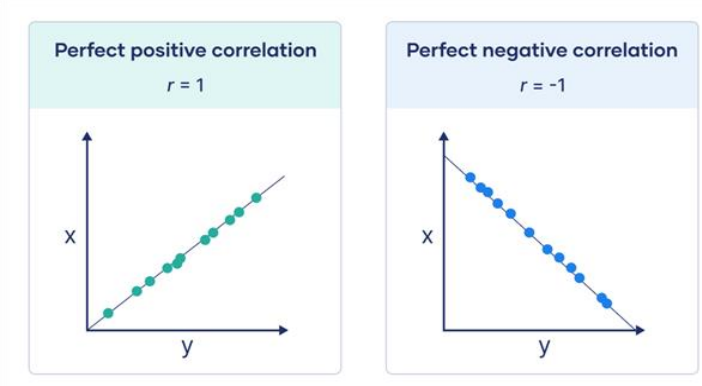
The Pearson coefficient gauges the strength and direction of the linear relationship between two continuous variables [62]. A continuous variable can take any numerical value in a given range or the entire range of real numbers. These variables may encompass an infinite range of potential values (e.g., a person's age, height, weight, or temperature). Put differently, the Pearson Coefficient delineates the extent of linear correlation between two quantitative variables, constituting a numerical-to-numerical evaluation. This coefficient ranges from -1 to 1. A value approaching 1 signifies a robust positive correlation, a value nearing -1 indicates a strong negative correlation, and a value near 0 suggests a weak or negligible correlation between variables (Table 3) [62]. A robust negative correlation signifies a tight and inversely proportional connection between two variables. Simply put, when one variable increases, the other consistently decreases. For example, if we consider the variable "outdoor temperature" and the variable "electricity consumption for heating", a strong negative correlation would mean that when the outdoor temperature increases, the electricity consumption for heating decreases significantly and vice versa.

Table 3
Interpretation of Pearson coefficient values [62]

Pearson correlation coefficient (*r*) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

In the context of generating synthetic data, evaluating Pearson correlation coefficient values can help determine how well the linear relationships between variables from the original data sets are preserved in the synthetic data [63]. If the values are consistent and similar, this increases confidence in the usefulness of the synthetic data in further analyzes and models.
To evaluate the preservation of relationships between variables in synthetic data using the Pearson correlation coefficient, a potential method involves calculating the Pearson coefficient for pairs of variables in both the original and synthetic datasets. Subsequently, a comparison of the Pearson correlation coefficient values for each variable pair between the original and synthetic data can be made. If the correlation coefficient values exhibit general similarity or proximity between the two datasets, it suggests that the linear relationships between the variables are maintained in the synthetic data.

Visual methods such as heat-maps can be used for this comparison [63]. If the heat map shows that the structure of the correlations is similar between the original and the synthetic data, this is a positive indication that the relationships between the variables are preserved. An alternative interpretation of the Pearson correlation coefficient (r) is as a gauge of how closely the observations align with a line of best fit [63]. Additionally, the Pearson correlation coefficient indicates whether the slope of the line of best fit is positive (Figure 9a) or negative (Figure 9b). In instances where the slope is negative, r takes a negative value, while in cases of a positive slope, r assumes a positive value. When r equals 1 or –1, all data points precisely lie on the line of best fit (Figure 9, a and b). Conversely, when r is 0, a line of best fit becomes ineffective in describing the relationship between the variables (Figure 10).



a ***b***

Figure 9: Line of best match

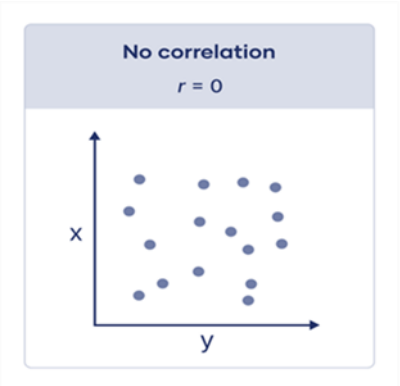


Figure 10: No correlation between variables

3.4.2.2 Cramer's V Correlation

Cramer's V is a measure of association or correlation used to assess the relationship between two nominal or dichotomous variables [64]. Nominal variables are variables that describe characteristics or attributes that can be grouped into discrete categories. These categories have no intrinsic order and are used to denote membership in a particular category or group. Examples of such variables are: Gender of a person (Male, Female, Other), Marital status (Married, Single, Divorced, Widowed), Favorite color (Red, Blue, Green, etc.). Dichotomous variables are a specific type of nominal variables that have exactly two possible categories or values. These can represent two-state attributes (e.g., a person's gender can be male or female), binary decisions or yes/no responses.
Cramer's V is based on the chi-square coefficient and provides a way to quantify the strength of association between nominal variables by considering the size of the contingency table (the cross-tabulation of the frequencies between the two variables) [64].

To calculate the Cramer's V coefficient, follow these steps [64]:

- The contingency table is calculated for the two nominal variables.
- Calculate the chi-square for the contingency table.
- The minimum number between the number of rows and the number of columns in the contingent table is determined. This number is denoted by "k".
- The Cramer's V coefficient is calculated using the formula:

  V = sqrt (X^2 / (n * k))

  Where:
  X is the calculated chi-square value.
  n is the total number of observations.

Cramer’s V correlation ranges from 0 to 1. A value nearing 0 suggests minimal association between variables, while a value close to 1 indicates a highly robust pairing (Table 4).

Table 4
Interpretation of Cramer's V coefficient values

Cramer's V	Strength
.25 or higher	Very strong relationship
.15 to .25	Strong relationship
.11 to .15	Moderate relationship
.06 to .10	Weak relationship
.01 to .05	None or negligible relationship

Cramer's V Correlation proves valuable in the context of appraising the effectiveness of synthetic data, particularly in assessing whether relationships between nominal variables are maintained in the generated synthetic data. This coefficient can provide information about the extent to which the distribution and associations between categories remain consistent between real and synthetic data. Assessing the association between nominal variables is important as it contributes to understanding how the synthetic data manage to capture the essential characteristics of the original data.

3.5 Evaluation of synthetic data privacy

In the increasingly digitized era of data processing, ensuring privacy is a particularly important concern in data use, especially when dealing with sensitive and personal data. In the context of generating synthetic data, privacy assessment plays a crucial role in ensuring that the resulting synthetic data can faithfully reflect the information in the original data without inadvertently revealing the identity or individual characteristics of the subjects. In this section, we will explore the data holdout-based evaluation method and discuss the use of statistical and detection evaluators to quantify the level of privacy and security provided by synthetic data.

3.5.1 Empirical evaluation based on synthetic data holdout

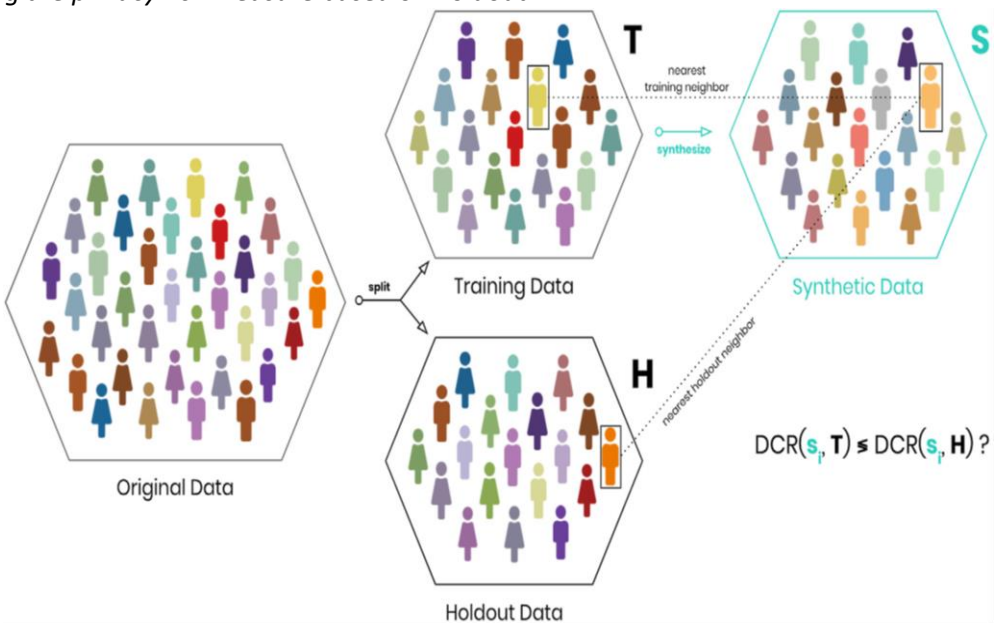
Although the generation of synthetic data has advanced in recent years, adequately evaluating the quality of these data remains an important challenge. The holdout data evaluation method serves as a crucial technique in statistical analysis for assessing the performance of a model or algorithm on fresh, unseen data [65]. This methodology entails dividing the dataset into two distinct parts: a training set, employed for constructing the model, and a holdout set, utilized to appraise the model's performance on novel data. The process consists of [65]:

- Data Split: The data set is split into a training set and a retention set with a common ratio such as 80-20 or 70-30. The larger set is used for training and the smaller set for evaluation.
- Model Training: The model or algorithm is trained on the training set to learn patterns from the data.
- Performance Evaluation: The model's performance is assessed on the retention set by generating predictions and comparing them to the actual values. Common metrics such as accuracy, precision, recall, and F1 score are employed to gauge the model's effectiveness.

The holdout method represents a nonparametric strategy devoid of explicit models and assumptions. It empirically gauges the faithfulness of a synthetic dataset to a designated target dataset by quantifying total variation distances (TVD) [66].

Figure 12

Constructing the privacy risk measure based on holdout



Note. From [66, p. 7]

For every synthetic record, we ascertain whether its nearest neighbor within the set of T training data is closer than the nearest neighbor within the holdout data H. The proportion of records exhibiting a closer proximity to training data than to holdout data becomes a metric for evaluating privacy risk. From the original dataset, two distinct sets are derived: a T training set, employed for synthetic generation, and an H holdout set, reserved without utilization by the generative model. The evaluation of privacy risk involves computing individual-level distances to the nearest record concerning the training data, as illustrated in Figure 12. Demonstrating that synthetic samples exhibit comparable proximity to both training and retained data provides robust evidence that the synthesizer has effectively learned to generalize patterns and operates independently of individual training records [66].

Total Variation Distance (TVD) is a key concept in probability theory and statistics, quantifying the dissimilarity between two probability distributions. It quantifies how far apart two distributions are from each other based on the differences in their probabilities assigned to various events or outcomes [67].

In mathematical terms, when P = (p_1, ..., p_x) and Q = (q_1, ..., q_x) represent two discrete probability distributions, the definition of the total variation distance for countable sets X is outlined as follows:

$$\delta(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)| .$$

Empirically, we should consider every possible event. Then, for each event, the probability assigned to both P and Q is calculated. For some events, the probability assigned to P will be higher, and for other events the probability assigned to Q will be higher. Then the entire list of events must be examined, and that event must be found for which the two probability assignments are the most different (it does not matter whether P or Q assigns the higher probability, it only matters that the difference between the assignments is maximal). The maximum gap between the assigned probabilities is referred to as the total variation distance. For instance, consider the scenario where a 6-sided die is rolled, and probability distribution P assigns a probability of 0 (zero) to the event "display the digit 1" while assigning a probability of 1/5 to the other five events. Further, suppose that Q assigns a probability of 1 to the event "display digit 1" and a probability of zero to all five other events. Each is a valid probability measure because each satisfies the

axioms. The "gap" between the assigned probabilities is 1 for the event "display digit 1" and 1/5 for the event "display digit 2" and 4/5 for the event "display digit 1 or 2", etc. Of the 6 possible events, it turns out that "display digit 1" has the largest offset, which is 1. This event is equal to the event "display digit 2, 3, 4, 5, or 6" which also has an offset of 1. So, the total variation distance is 1 (one).
In the context of evaluating synthetic data quality, Total Variation Distance is often used to compare the distributions of synthetic and original datasets. It helps assess how well the synthetic data captures the distributional characteristics of the original data. The process involves calculating the TVD between the distributions of individual variables in the two datasets [65]. A lower TVD value indicates a closer match between the distributions, suggesting that the synthetic data is more faithful to the original data.
This method, together with other evaluators and metrics, contributes to the complete assessment of disclosure risk and the usefulness of synthetic data, providing crucial information for decision-making and the safe and effective use of this data. Next, we will describe the evaluators and metrics used in this research and show their contribution to the evaluation of the confidentiality of synthetic data.

3.5.2 Statistical Evaluator

Statistical metrics are designed to gauge the level of concordance between the statistical attributes of the synthetic data and those of the original dataset. This examination focuses on appraising how well the distribution, structure, and features of the synthetic data align with those of the authentic dataset. Significant disparities in statistical properties may indicate a restricted usefulness of the synthetic data. Simultaneously, excessive similarity could pose privacy risks and undermine diversity [68].

3.5.2.1 Kolmogorv-Smirnov Test

Definition [68]: The Kolmogorov-Smirnov (KS) test is a non-parametric technique employed to determine if two samples originate from identical distributions. It evaluates the likeness of distributions by comparing their empirical distribution functions (ECDF). Given a set of N ordered data points Y1, Y2, ..., YN, the ECDF is defined by the formula:

$$\mathbf{E_N = n(i)/N}$$

Where: n(i) is the number of points less than Y_i and Y_i are ordered from the smallest to the highest value.

Algorithm [69]:

- The Kolmogorov-Smirnov test establishes two hypotheses: Null hypothesis (H0) - the data come from a specified distribution (both samples were drawn from populations with identical distributions) and alternative hypothesis (H1) - at least one value does not fit the specified distribution.
- The KS test calculates the D statistic that represents the largest vertical difference between the two ECDFs. This statistic is subsequently compared to the critical values from the Kolmogorov distribution to assess whether a significant difference exists between the distributions. The Kolmogorov-Smirnov test statistic, denoted D, is defined as:

$$\mathbf{D = \max_{1 \le i \le N} (F(Yi) - (i-1)/N,\ i/N - F(Yi))}$$

  Where: F represents the cumulative theoretical distribution of the examined distribution, which must be continuous and thoroughly specified.

Values and Interpretation [69]: The KS test highlights the maximum discrepancy between the cumulative distributions of the two samples and computes a P-value based on this difference and their respective sample sizes. As the test does not compare a specific parameter (such as mean or median), it does not provide a confidence interval. The interpretation of the P-value revolves around the question "What is the probability that the Komogorov-Smirnov D statistic's value is as large or larger than the observed one?". A

small P-value suggests that the two samples originate from populations with distinct distributions. Conversely, a larger P-value indicates potential similarity in the distributions.
When assessing the confidentiality of synthetic data [70], the interpretation of test results is as follows: A low P-value indicates dissimilar distributions between the synthetic and original data, suggesting a potential privacy risk. If the P-value is large, this suggests that the distributions are similar, which could indicate better privacy protection. To interpret the P-value, a privacy threshold can be set. For example, we can decide that P-values less than 0.05 indicate significant privacy exposure, while P-values greater than 0.05 indicate better privacy protection. It is important to note that the Kolmogorov-Smirnov test provides a measure of the difference between the distributions but does not provide detailed information about the nature or source of the difference. As such, it represents just one of the approaches available for evaluating the privacy of synthetic data and should be employed alongside other assessments and measures to ensure a comprehensive privacy evaluation.

3.5.2.2 Chi-Squared Test

Definition: The Chi-Square test is a statistical method that measures the degree of discrepancy between observed frequencies and expected frequencies in a contingency table consisting of two categorical variables [71].

Algorithm [72]:

- The chi-square test is defined for two hypotheses: The null hypothesis (H0) posits that the data adhere to a predefined distribution, while the alternative hypothesis (H1) suggests that the data deviate from the specified distribution.
- A contingency table is constructed that frames the categorical variables to be tested.
- Determine the expected frequencies for each cell in the table, relying on the marginal distribution of the two variables.
- The Chi-Square statistic is calculated using the formula:

$$\chi^2 = \Sigma \left((\text{observed frequency} - \text{expected frequency})^2 / \text{expected frequency}\right),$$

  where the sum is done over all cells of the table.
- The computed Chi-Square statistic is juxtaposed with a critical value obtained from the Chi-Square distribution, considering a specific degree of freedom and significance level.

Values and Interpretation [73]: The values obtained from the Chi-Square test are compared with the critical values in the Chi-Square table corresponding to a certain predetermined level of significance. If the calculated Chi-Square test value surpasses the critical value, it enables us to reject the null hypothesis, indicating a substantial association between the two examined variables. Although, Chi-Square tests provide a measure of the relationship between variables, these values alone do not provide a complete understanding. For example, we cannot determine whether a Chi-Square value of 1.3 indicates a poor or strong fit. To correctly interpret the results of the Chi-Square test, it is necessary to compare the obtained statistics with the corresponding Chi-Square distribution. This allows us to decide whether the null hypothesis should be rejected or not. A crucial element in utilizing the test involves determining a particular level of significance, commonly denoted as α, which signifies the probability of erroneously rejecting the null hypothesis. For example, at a confidence level of 95% (or $\alpha = 0.05$), there is a 5% risk of making an error and incorrectly rejecting the null hypothesis.
In other words, if:
• A p-value of ≤ 0.05 indicates substantial evidence against the null hypothesis, leading to the conclusion that the data deviate from a distribution with specific proportions.
• Conversely, a p-value > 0.05 suggests insufficient evidence to reject the null hypothesis, meaning we cannot conclusively state that the data adhere to the distribution with specified proportions. However, it is important to note that this does not imply the distributions are identical; there might be differences, but the test lacks the power to detect them.

In the evaluation of the privacy implications of synthetic data, the Chi-Square Test can be employed to ascertain whether the distributions of categories in the synthetic data closely align with those in the original data. One can calculate the Chi-Square test for the corresponding categorical variables between the two data sets and compare the results to see if there are significant differences between the distributions [74].

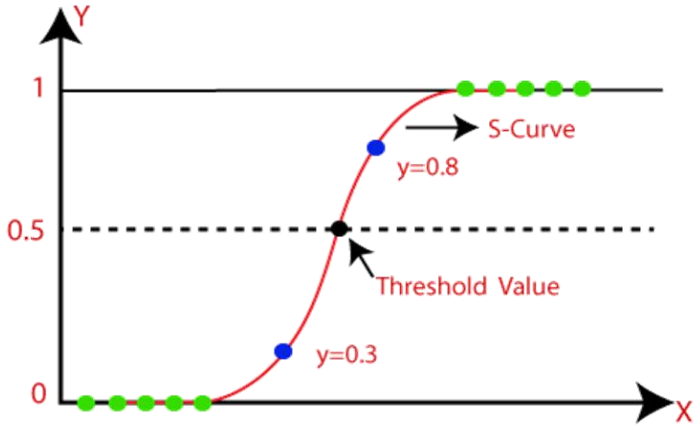
3.5.3 Data Detection Evaluator

Data detection is an effective method used to assess the privacy of synthetic data, involving advanced analysis techniques to identify and quantify the risk of exposure of sensitive information in the generated datasets. This approach is based on two key techniques: Logistic Regression and the Random Forest Classifier Algorithm, which are applied to examine and classify synthetic data according to the risk associated with the disclosure of confidential information [75].

3.5.3.1. Logistic regression

Definition: Logistic Regression is a statistical method employed to model the association between a binary dependent variable and one or more independent variables. This technique is particularly useful when we want to understand and predict the probability of an event with two possible outcomes. As a result, the output is inherently categorical or discrete. It may manifest as "Yes" or "No," 0 or 1, "True" or "False," etc. However, instead of providing precise values like 0 and 1, it yields probabilistic values within the range of 0 to 1 [76]. In logistic regression, the methodology deviates from fitting a linear regression line, as observed in linear regression, to fitting an "S"-shaped logistic function capable of predicting two potential outcomes (0 or 1). The curve derived from the logistic function represents the probability, specifically in our context, of whether the examined data is genuine or not. Logistic regression demonstrates versatility in classifying observations across diverse data types and adeptly identifies the most influential variables for classification purposes, as delineated in reference [77]. Figure 13 shows the logistic function.

Figure 13
Logistic function (sigmoid function)



Note. From [78]

Algorithm: The sigmoid function, highlighted in reference [79], serves as a mathematical tool for transforming predicted values into probabilities. It operates by mapping any real value to another within the range of 0 and 1. Notably, logistic regression necessitates that its values fall within the 0 to 1 range, forming a distinctive "S" shaped curve known as a sigmoid function. Within logistic regression, the notion of a threshold value is integral, defining the probability of either 0 or 1. Consequently, values surpassing the threshold lean towards 1, while those below the threshold gravitate towards 0. In *Figure 13*, the y-axis values range from 0 to 1, intersecting the axis at 0.5.
The equation for logistic regression is [80]:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

Where:

- p represents the probability of the event taking place
- x_1, x_2, ……., x_n are the predictor (independent) variables.
- b_0, b_1, b_2, …, b_n are the coefficients estimated by the logistic regression model.

This equation represents the natural logarithm of the odds ratio of the event occurring (probability of success) to the event not occurring (probability of failure), and it is transformed linearly with the predictor variables. The logistic regression model estimates the coefficients (b) to best fit the observed data, which allows for the prediction and inference of the probability of the binary outcome based on the predictor variables. The coefficients offer insights into the direction and magnitude of the relationship between the predictors and the log-odds of the event [80].

Categorized by types, logistic regression can be delineated into three distinct forms [82]:

- ✓ Binomial: Binomial logistic regression involves scenarios where there are only two possible types of dependent variables, such as 0 or 1, Pass or Fail, etc.
- ✓ Multinomial: In multinomial logistic regression, there exist three or more possible unordered types for the dependent variable, such as "cat", "dog", or "sheep";
- ✓ Ordinal: Ordinal logistic regression encompasses situations where there are three or more possible ordered types of dependent variables, such as "Low," "Medium," or "High".

When implementing logistic regression, it is essential to consider the following assumptions [82]:

- Binary logistic regression necessitates binary target variables, focusing on predicting outcomes for the level of factor 1;
- The model should avoid multicollinearity, ensuring independence among its variables;
- Significance is crucial; thus, the model must include variables that significantly contribute to the predictive power;
- Logistic regression requires a substantial sample size to yield reliable results.

Values and Interpretation: The coefficient values reflect the direction and magnitude of the influence that each independent variable has on exposure risk. These values are presented in the form of log-odds and can be interpreted as follows: an increase in the coefficient by one unit corresponds to a proportional increase in the log-odds and, implicitly, the probability. It is essential to emphasize that logistic regression relies on the assumption of a linear relationship between the log-odds of the outcome and the predictor variables. This is why the equation incorporates the natural logarithm. The logit transformation is employed to ensure that the predicted probabilities fall within the bounds of 0 and 1, making it appropriate for modeling binary outcomes [81].

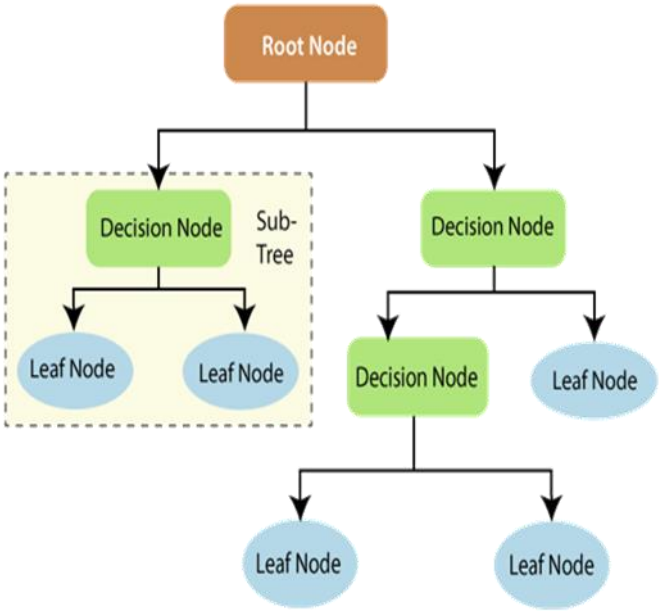
Logistic regression is a valuable method in assessing the privacy of synthetic data [80], providing detailed understanding of the risk of exposure and the factors that influence it. By using logistic regression, we can analyze how the predictor variables are associated with the probability of exposure within the synthetic data. This allows us to identify which of these variables have a significant influence on exposure risk and to quantify this influence. Interpreting logistic regression results involves evaluating the estimated coefficients (b) for each predictor variable. These coefficients reflect how changes in the predictor variables are associated with changes in the log-odds of exposure risk. A positive coefficient signifies a rise in log-odds (and, consequently, an increase in exposure risk) as the predictor variable's value increases. Conversely, a negative coefficient implies a decline in log-odds and exposure risk as the value of the predictor variable increases. The p-values linked to the coefficients are employed to evaluate the statistical significance of the identified relationships. If the p-value falls below the selected significance level, typically 0.05, it leads to the conclusion that there exists a significant relationship between the predictor variable and the exposure risk. Therefore, logistic regression allows us to examine in depth the impact of each variable on the privacy of synthetic data and make informed decisions regarding privacy protection measures.

3.5.3.2. Random Forest Classifier

Definition: Random Forest Classifier is a powerful and versatile technique in synthetic data privacy assessment, bringing to the fore a complex understanding of exposure risk and major influences. This approach is based on the construction of an ensemble of decision trees, which operate together to make accurate predictions and to identify the importance of different variables on exposure risk [83]. By implementing the Random Forest Classifier, we can assess how predictor variables interact to influence exposure probability in synthetic data. This analysis allows us to uncover the complex combinations of factors that contribute to exposure risk and to quantify the contribution of each variable to that risk.
Algorithm: A random forest constitutes a supervised machine learning algorithm constructed based on decision tree algorithms. The Random Forest Classifier algorithm assumes [84]:

- Initialization: Starts with a training dataset containing labeled examples, where each example consists of a set of features (variables) and a class label.
- Construction of Decision Trees:
  - A random subset of the training set, called the "bagging set", is chosen.
  - A decision tree is constructed on this subset using a construction algorithm such as the Classification and Regression Trees (CART) algorithm. A decision tree is a decision support method that creates a structure resembling a tree. The three elements of the decision tree, namely decision nodes, leaf nodes, and the root node, are depicted in Figure 14. A decision tree algorithm partitions a training dataset into branches, which then further subdivide into additional branches. This process continues until a leaf node is reached, and a leaf node cannot undergo further splitting. The final output generated by each specific decision tree is represented by its respective leaf node.

Figure 14
The three types of nodes in a decision tree



Note. From [86]

  - Each decision tree is built with a limit on the depth and/or the maximum number of features to avoid overfitting.
- Generation of the Ensemble:
  - Step 2 is repeated several times to build a predetermined number of decision trees.
  - These decision trees form the Random Forest ensemble.
- Predictions:
  - For each example from the test data set or from the synthetic data, the example is passed through each tree in the assembly.

- Each tree outputs a class prediction for the instance.
- The ultimate prediction is established through a majority vote in the case of classification or by averaging in the case of regression, considering the predictions generated by all the trees in the ensemble. In this scenario, the outcome favored by the majority of decision trees serves as the ultimate result of the random forest system. For instance, if four trees predict the authenticity of the data and three trees predict otherwise, the final prediction will be that the data is authentic.

- Rating and Importance of Features:
  - Evaluate ensemble performance by measuring accuracy, recall, precision, or other relevant metrics.
  - Feature importance is calculated by analyzing how much each feature contributes to improving the performance of the ensemble.

Values and Interpretation [85]:

- Accuracy, Precision, Recall: These are common evaluation metrics used to measure the performance of the Random Forest Classifier on both real and synthetic data. High accuracy indicates that the classifier is able to correctly predict class labels, while high precision and recall indicate that the classifier is effectively identifying positive cases (sensitive data) and minimizing false positives and false negatives. If the Random Forest Classifier exhibits similar performance (e.g., accuracy, precision, recall) on both real and synthetic data, it suggests that the synthetic data captures the underlying patterns of the original data without revealing additional sensitive information.
- Scores of Feature Importance: Random Forest assigns importance scores to features, indicating their contribution to the classifier's accuracy. In the privacy assessment context, higher scores for certain features may indicate potential vulnerabilities where synthetic data could unintentionally disclose sensitive information.
- Out-of-Bag (OOB) Error Rate: Random Forest leverages OOB samples (data not utilized during training) to gauge the classifier's error rate. If the OOB error rate is noticeably higher for synthetic data in contrast to real data, it could indicate that the synthetic data is less secure and potentially reveals sensitive patterns.
- Confusion matrix: Through an examination of the confusion matrix, one can discern the true positives, true negatives, false positives, and false negatives of the classifier, considering both real and synthetic data. Analyzing the confusion matrix can provide insights into the types of misclassifications made by the classifier on synthetic data. If certain classes are consistently misclassified or show different patterns compared to real data, it may suggest potential privacy risks.

In the evaluation of the privacy aspects of synthetic data [87], the Random Forest Classifier can be employed through the following steps:

- Model Training: A Random Forest Classifier is developed using the original (non-synthetic) data to construct the classification models for the real data.
- Predictions on Synthetic Data: Use the trained model to make predictions on synthetic data. If the classification models correctly predict the class labels from the synthetic data, this can indicate a possible exposure of confidential information.
- Analysis of Feature Importance: We examine the significance of features in the Random Forest model to identify those with a substantial impact on predictions in the synthetic data. Features of high importance may indicate potential exposure risks.

3.5.4 Duplicate Evaluator

The Duplicate Evaluator is a component of the synthetic data evaluation process, with the role of identifying and quantifying the degree of similarity or overlap between synthetic and real records from an original data set [88]. This evaluator looks for synthetic records that are identical or very close to real records, highlighting potential privacy risks or disclosure of sensitive information from the synthetic data.

3.5.4.1 Gower Evaluator

One of the methods used to evaluate duplicates is the ***Gower Evaluator***, which relies on Gower distances to quantify the similarity between records. The Gower Evaluator method uses Gower distances to calculate the similarity between the records in the original dataset and the synthetic ones. This technique considers the mixed nature of the data, including continuous and nominal variables. Gower distances consider absolute differences for continuous variables and similarity measures for nominal variables [89].

Algorithm: Gower's methodology relies on a series of steps to determine a measure of similarity or distance between two records. In summary, these steps consist of [90]:

- Normalization (if applicable): For continuous or ordinal variables, normalization to the interval [0, 1] may be required to ensure that all variables have the same range of values.
- Calculation of distances for continuous and ordinal variables:
  - For continuous and ordinal variables, the absolute difference or the squared difference (depending on preference) between the corresponding values of the records is calculated.
  - The calculated distances are standardized to the interval [0, 1] by dividing them by the range of possible values of the variable.
- Evaluating distances for nominal variables involves assigning a distance of 0 when the values are identical and 1 when they differ.
- Calculation of Gower distances:
  - For each pair of records, the distances for each variable are calculated using the methods above.
  - The weighted average of these distances is calculated to obtain the Gower distance between records.

Usage of Gower distances:

- Gower distances can be used to identify the most similar or most dissimilar records in the data set.
- Having a way to calculate the distance between two individuals (Gower distance), we can use it to build privacy measures (DCR and NNDR).

3.5.4.2 Distance to Closest Record (DCR)

Definition and Algorithm [91]:
Distance to Closest Record (DCR) is a measure used in evaluating synthetic data against real data. This metric centers on the proximity between each synthetic record and the nearest authentic record within the source dataset. The objective is to evaluate the extent to which the synthetic data replicates the original dataset and to gauge the potential risk of disclosing sensitive information. We can define the distance to the nearest record for a given individual, denoted by DCR(s) for individual "s" in the synthetic data set "S", as the minimum distance between "s" and each original individual "o" in the set of original data "O":

$$\textbf{DCR(s) = min d(s,o) for each o}\in\textbf{O}$$

This metric expresses how close the synthetic individual "s" is to the nearest original record "o" in terms of the specified distance, where "d" represents the distance function used (e.g., Gower distance).

Values and Interpretation [89]:
DCR values can vary from 0 to a certain maximum, depending on the scale of the distance metric used. DCR(s) = 0 indicates that the synthetic instances "s" is an identical copy (clone) of at least one real instances in the original data set "O". The higher the value of DCR(s), the more distant the synthetic instances "s" is from the original records, suggesting a lower similarity to the real data and, implicitly, a lower risk of privacy violation.
Additionally, the $5th$ percentile of this value (P5) is calculated for the DCR(s) values for all synthetic instances in the synthetic data set. This represents the DCR(s) value below which 5% of the synthetic instances lie. The choice of this percentile is important because we are interested in identifying the highest

privacy risks, and P5 provides a robust estimate for this. Essentially, the higher the P5, the less synthetic instances are close to the original records, indicating a lower privacy risk. This value can be used to make informed decisions about the use of synthetic data and to ensure that the risk of disclosure of sensitive information is kept under control.
In addition to evaluating the DCR between the generated synthetic data set and the original data set, according to the previously described method, an additional approach is to calculate the DCR between the synthetic data set and a holdout data set [92]. This approach aims to ensure that synthetic individuals are not systematically closer to individuals in the original data set than to those in the holdout set. By calculating the DCR values for each synthetic instance against both the original set and the holdout set, we can determine the percentage of synthetic instances that are more similar to instances in the original set than those in the holdout set. The goal is to get a percentage as close to (or below) 50% as possible. This indicates that the synthetic data does not contain meaningful information that could allow an attacker to infer whether a particular person was present in the real data set [92]. Therefore, a robust assessment of the risk of privacy disclosure in the context of synthetic data is obtained.

3.5.4.3 Nearest neighbor distance ratio (NNDR)

Definition: Neighbor Distance Ratio (NNDR), or nearest neighbor distance ratio, is a measure that tells us how the points in a data set are arranged. It helps us to understand if the points are grouped together in a certain way or if they are evenly distributed in space [93].

Algorithm: To assess the confidentiality of synthetic data using the NNDR (Nearest Neighbor Distance Ratio), a potential algorithm involves the following steps [94]:

- Calculation of NNDR for Real Data:
  - For each data point, determine the distance to the nearest neighbor and the second nearest neighbor using the robust Gower methodology.
  - Calculate the ratio between the distance to the nearest neighbor and the distance to the second nearest neighbor for each point.
  - Compute the average of these ratios (NNDR) using the formula:

NNDR = (Σ (distance to nearest neighbor / distance to second nearest neighbor)) / total number of points

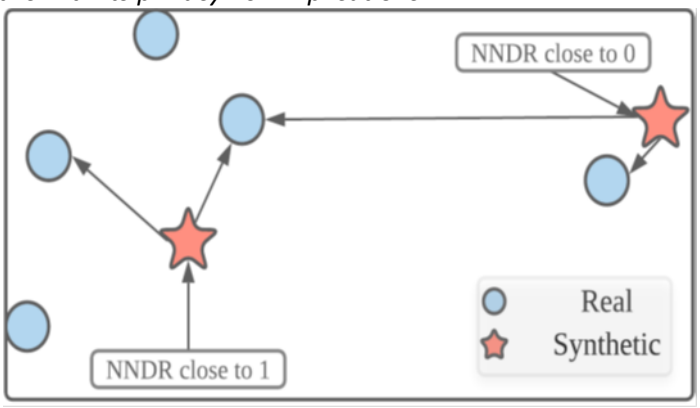
Where: Σ represents the sum, and distances are as defined above.

- Calculation of NNDR for Holdout: Repeat the same steps as described above for the holdout data set. Calculate the distance to the nearest neighbor and the distance to the second nearest neighbor, then determine the ratios and their average for the holdout data set.
- Comparison of NNDR between Synthetic and Real: Utilize the same steps to calculate the NNDR for the synthetic data set. Compare the NNDR values obtained for the synthetic and real data sets to evaluate the level of confidentiality achieved by the synthetic data.

Interpretation of Results: The interpretation is based on the average value resulting from the calculations of the NNDR reports for the entire synthetic data set [95]. Low values of NNDR (approaching 0) suggest that the majority of points in the synthetic dataset are concentrated or clustered near the points in the real dataset. This suggests better privacy because synthetic data mimics the distribution and structure of points in real data. Large values of NNDR (close to 1) suggests that the points in the synthetic data set are evenly distributed and do not show a tight cluster around the points in the real data. This may indicate an increased risk of sensitive data exposure, as synthetic data fails to reflect the natural clustering of real data (Figure 15). In essence, the interpretation of the NNDR average focuses on evaluating whether the synthetic data can reproduce the distribution and structure of the real data in terms of the neighborhood of the points. A low value of NNDR suggests that the synthetic data closely approximates its real neighbors, which enhances privacy, while a high value indicates a possible deficiency in privacy protection.

Figure 15

Illustration of NNDR metric with its privacy risk implications



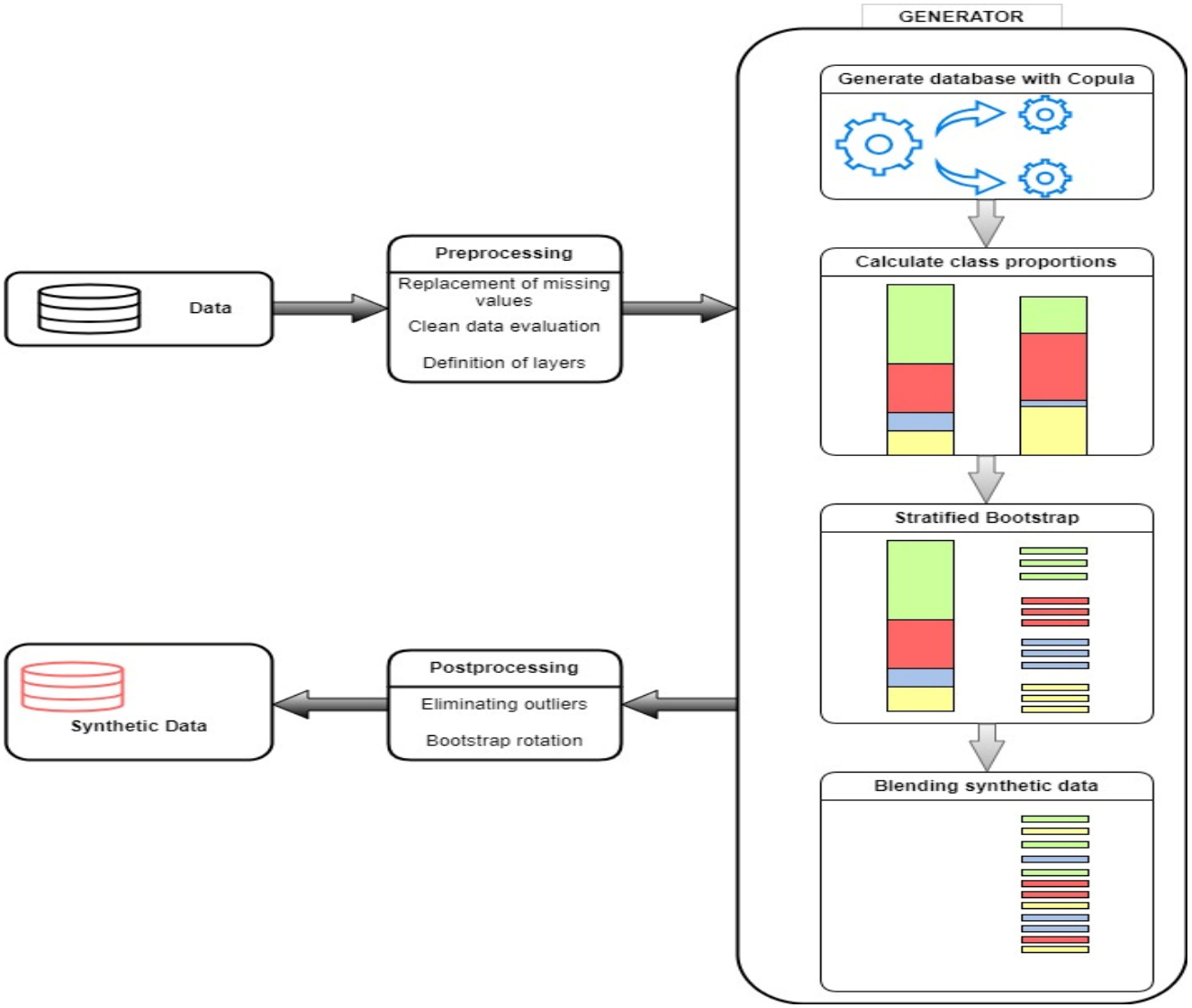
Note. From [95, p. 10]

If the calculated NNDR values for the synthetic data are compared to the NNDR values for the real and holdout data [94], a small NNDR value for the synthetic data relative to the real and holdout indicates that the synthetic data is distributed similarly to the real data and does not expose sensitive information. If the NNDR value for the synthetic data is closer to the NNDR value for the holdout data than to the NNDR value for the real data, it indicates that the synthetic data preserves privacy better than the real data. In this way, the adapted algorithm will provide a more comprehensive assessment of the privacy of the synthetic data by comparing the NNDR between the synthetic, real, and withheld data and by analyzing how the synthetic data is distributed relative to both reference data sets.

Chapter 4
The “Fusionstrap” Method

In this thesis, we propose the “Fusionstrap” framework, an integrated data processing and synthesis system, developed as a response to the challenges of class imbalances in datasets. This framework addresses multiple essential aspects of data processing, including rigorous preprocessing, advanced synthetic data generation using layered Bootstrap to address class imbalances, and post-processing techniques such as outlier removal and applying Bootstrap rotation to obtain a unique set of synthetic data. Comprehensive evaluation of the utility and privacy of synthetic data, compared to other data synthesis methods, completes this framework. The diagram of the relevant components (Figure 16) provides a visual illustration of the architecture and functionality of the “Fusionstrap” system, showing how each stage contributes to improving data quality and ensuring confidentiality, in a context of effective class imbalance management [19].

Figure 16
Method visualization



The structure of the "Fusionstrap" method is divided into three distinct components. In the first phase, the preprocessing of the data sets is carried out, which includes the replacement of missing values, the evaluation of the cleaned set and the definition of the layers with the calculation of the percentages of class imbalances (Section 4.1). The processed data subsequently proceeds to the primary stage of the approach, involving the generation of 150 synthetic datasets (Section 4.2). Ultimately, the method

incorporates a post-processing step aimed at maintaining the coherence of the synthetic data and enhancing its variability (Section 4.3).

4.1 Data Preprocessing

The data preprocessing segment holds a pivotal role in upholding the quality and coherence of the data utilized by the "Fusionstrap" framework. In the initial phase of this procedure, the resolution of missing values in the dataset is undertaken to mitigate potential distortions or ambiguities that might impact the subsequent analysis and synthesis of the data. The cleaned data set is evaluated in the next step to determine to what extent the characteristics of the initial set have been affected by the cleaning. The preprocessing ends with the definition of the layers for each variable and the calculation of the percentages of class imbalances.

4.1.1 Replacement of missing values

Depending on the specifics of each variable, we opted for the following replacement techniques:

- **Missing numerical variables** are replaced by the arithmetic mean of the known values (Mean Imputation) or by the median value of that variable (Median Imputation) [96]. The choice between using the mean and the median to replace missing values in numerical variables depends on the distribution of the data and possible outliers [96]. If the data are approximately symmetrical and there are no significant outliers, then the missing values are replaced by the mean. The mean uses all available values for the calculation and is more sensitive to small variations in values, which can better reflect the overall trend of the data set. On the other hand, if the data have a skewed distribution or there are significant outliers, the median is used to replace the missing values. The median represents the central value of the data set when the values are ordered ascending or descending and is more robust to extreme values. Using the median can avoid the influence of outliers and provide a more stable estimate of the central tendency of the data, reducing the risk of significant dispersion distortion.
- **For non-numeric variables** such as categories or modalities, we opted to replace with the most frequent category or modality in that variable (Mode imputation) [96]. This helps maintain the structure of the categories and preserve the relative proportions between them.

4.1.2 Clean data evaluation

Furthermore, to assess whether statistics and relationships between variables have been preserved in the preprocessed data set, the “Fusionstrap” framework performs a comparative analysis between the preprocessed data and the original data. This analysis focuses on the following key aspects:

- **Checking the basic statistics of the cleaned data**: the basic statistics of the variables (such as mean, median, standard deviation, minimum and maximum) are compared between the original and the preprocessed dataset. This comparison provides insight into how data cleaning affected the underlying statistics. Examining these differences can assess the impact of the cleanup and provide assurance that the changes are in line with expectations. If these statistics are preserved properly, it means that the preprocessing did not introduce significant distortions in the data distribution. Also, if differences occur that could significantly affect the analysis, the data cleaning process should be re-evaluated and alternative methods applied or we should investigate in more detail the source of differences [97].

- **Checking the correlations analysis of the cleaned data set:** The correlation matrix between the variables is calculated and the correlations before and after preprocessing are compared. To calculate the correlation matrix, “Fusionstrap” uses a combination of Pearson's correlation coefficient for numerical variables and Cramer's V correlation coefficient for nominal variables. The correlation matrix is made with Bias Correction. The preference for Cramer's V with Bias Correction is motivated by the fact that this method provides more accurate and robust results compared to the classic Cramer's V method. This is because Cramer's V with Bias Correction takes outliers into

account and provides a more accurate estimate of the correlation. By adjusting the expected values to account for the outliers, the Bias Correction method removes some of the influence on the correlation that can only be attributed to differences in the outliers. Therefore, in general, the results obtained with Cramer's V method with Bias Correction are considered better and more accurate compared to the classical method. This method is particularly useful when working with nominal or ordered variables and you want a more accurate estimate of the correlations between them [97].

Overall, evaluating the retention of statistics and relationships between variables in the preprocessed set involves a detailed and comparative analysis between the original and processed data. If it is observed that the changes made by the preprocessing are minimal and that the essential structures and relationships are preserved, then we can be confident in the quality of the data preprocessing. If, however, significant changes are found in the pre-processed data, it is crucial to identify and address these discrepancies. This involves identifying the variables that have undergone significant changes and determining whether these changes are consistent with expectations. It may also be necessary to adjust parameters or explore other preprocessing techniques to achieve more consistent results.

4.1.3 Definition of layers

The original data set is divided into several layers based on the existing classes. Layers are defined based on the classification variable (e.g., outcome variable or label). Each layer represents a distinct category of the classification variable, such as a majority class and a minority class. For each class, a probability distribution is constructed based on the frequency of each layer.

The “Fusionstrap” framework performs layer definition and analysis through the following steps [25]:

- Initial data cleaning and preparation: It starts with loading the initial data set. The categorical columns that will be used to generate synthetic data are identified. The identification of the categorical columns that will be used to generate the synthetic data is done by traversing the predefined list of column names from the initial set. These columns are then converted to columns of data type 'str', thus ensuring that all values are treated as strings in the subsequent synthetic data generation process.
- Selection of categorical data for analysis: Only categorical variables in the data sets are selected to calculate the class distribution and class imbalance for them. To calculate the class distribution and class imbalance, “Fusionstrap” defines two functions: the first function receives a data frame and calculates the class distribution for each variable, and the second function receives the previously calculated class distribution and calculates the percentage of class imbalance for each variable. To calculate the percentage of class imbalance for each variable, the following formula is used:

  (max_count - min_count) / max_count * 100,

  Where: max_count represents the maximum number of instances in a class, and min_count represents the minimum number of instances in a class. This provides a clearer understanding of how classes are distributed within each variable.
- Display of results: class distribution and percentage of class imbalance are shown for each variable. Interpreting class imbalance percentages gives us insight into the degree of inequity in the class distribution for each variable. Variables with high percentages of imbalance indicate significant inequity between classes.

Defining layers is only an initial step towards assessing and improving data quality and addressing issues of class imbalance in the dataset. If we take the example of the Diabetes Prediction data set [104], we will have the classification variable "Diabetes" with two classes: "Positive" and "Negative". Based on this classification variable, the data set will be divided into two layers: the "Diabetes Positive" layer, which includes data associated with people diagnosed positively with diabetes and the "Diabetes Negative" layer, which includes data associated with people diagnosed negatively for diabetes. Depending on the probability distributions built on the basis of frequency for each of these two classes, we will be able to determine which of them is the majority and which is the minority.

By analyzing each layer in detail, we can better understand the characteristics associated with people who have been diagnosed positive for diabetes and ensure that the synthetic data generation algorithm considers the specificities of both classes in its process. This layered approach is useful when we want to focus on improving the representation of the minority class (for example, people with positive diabetes) to avoid the problems generated by class imbalances.

4.2 Synthetic data generation

“Fusionstrap” generates synthetic data by using a hybrid algorithm based on Stratified Bootstrap and Gaussian Copula Synthesizer. This approach combines the technique of Stratified Bootstrap that ensures adequate representation of each class with the power and flexibility of the Gaussian Copula Synthesizer [99] that preserves the consistency and reality of the generated data. The goal is to produce reliable synthetic data that not only reflects the original nature of the data set, but also mitigates inequities between classes for more robust and relevant analytical results. The hybrid algorithm "Stratified Bootstrap with Gaussian Copula Synthesizer" consists of generating the synthetic database with the Gaussian Copula Synthesizer, calculation of class proportions, generating Bootstrap stratified samples and blending synthetic data samples.

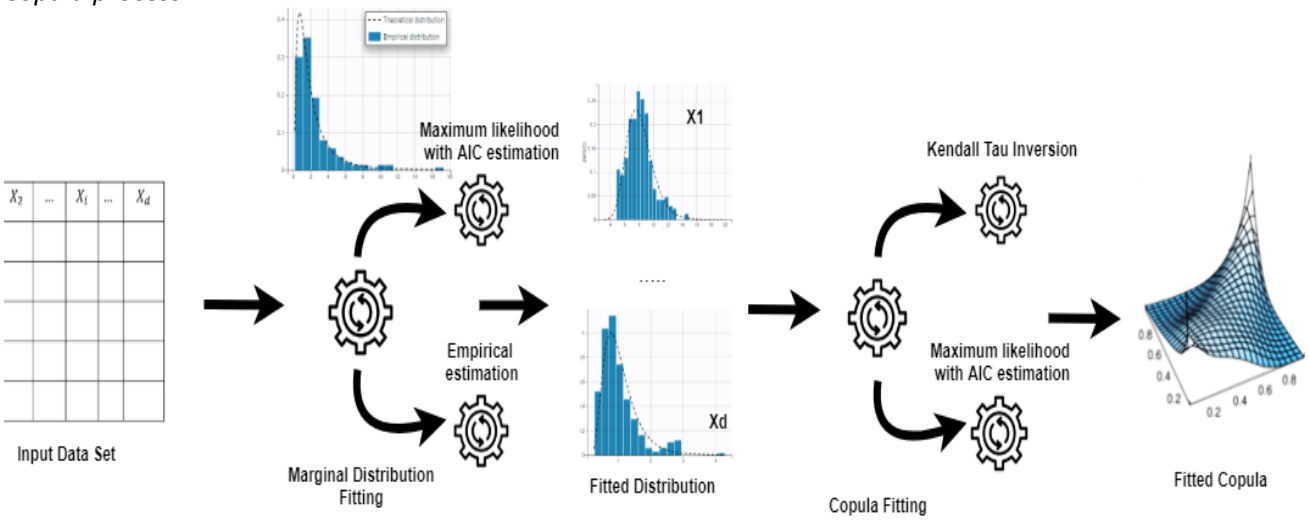
4.2.1 Generating the synthetic database with the Gaussian Copula Synthesizer

The algorithm of Gaussian Copula Synthesizer [99] assumes:

- For each class in the original data, a Gaussian Copula Synthesizer model is built based on the metadata of the original data set. In this step, a technique called "Gaussian Copula Synthesizer" is used to model the relationships between the variables in the original data set. This is a synthesis algorithm based on Gaussian distributions and copulas, which are mathematical functions used to capture complex dependencies between variables. To illustrate, suppose we have two variables, "education" and "income," and we want to generate synthetic data based on the distributions and relationships between them (Figure 17).
- The Gaussian Copula Synthesizer model is fitted to the data of the specified class. Fitting means that the algorithm learns the specific distributions and relationships of the data for this class so that it can generate synthetic data that preserves these characteristics. Example: For the class labeled "0" in the original data set, the model learns how the variables "education" and "income" are distributed for this class specifically (Figure 17).
- A synthetic database is generated using the fitted model, with size based on the total number of Bootstrap samples: The Gaussian Copula Synthesizer model is applied to create synthetic data for the present class. The number of Bootstrap samples is used to decide how many synthetic data to generate for this class so that the correct proportions between classes are preserved. Example: If we have 100 Bootstrap samples and class "0" represents 30% of the original data, then approximately 30 Bootstrap samples will be generated for this class. These samples will preserve the previously learned distributions and relationships for the "education" and "income" variables.

Figure 17

Copula process



Note. From [99, p. 6]

4.2.2 Calculation of class proportions

After the data for each class has been generated using the Gaussian Copula Synthesizer, the proportions of each class in the synthetic database are calculated. This stage is crucial to guarantee the preservation of the initial class distribution in the synthetic data, ensuring an accurate reflection of the class imbalances present in the original dataset.

Continuing the previous example with two classes and a total of 150 Bootstrap samples, in the set of initial data, we have the following class distribution:

Class "0": 30% (about 45 samples out of a total of 150)

Class "1": 70% (about 105 samples out of a total of 150)

After generating the synthetic data for each class using the Gaussian Copula Synthesizer and obtaining the corresponding Bootstrap samples, suppose the results are as follows:

For class "0", we generated about 50 Bootstrap samples (the value may vary due to the random generation process).

For class "1", we generated about 100 Bootstrap samples (the value may vary due to the random generation process).

The next step is to calculate the class proportions in the synthetic database: For class "0", the number of samples generated is 50. Percentage-wise, this represents about 33.33% of the total of 150 samples in the synthetic database. For class "1", the number of generated samples is 100. Percentage-wise, this represents approximately 66.67% of the total 150 samples in the synthetic database. Thus, the class proportions in the synthetic database would be around 33.33% for class "0" and 66.67% for class "1". These proportions reflect the original distribution of classes in the original data set and ensure that class imbalances are preserved in the generated synthetic data.

4.2.3 Generating Bootstrap Samples

For every Bootstrap sample, an empty data frame is generated specifically for that sample. This frame will be utilized to incorporate samples from each class individually. The number of samples for each class within the Bootstrap sample is then computed based on the pre-determined class proportions. In the example above, for class "0" we have approximately 33.33% of 150, i.e., 50 samples, and for class "1" we have approximately 66.67% of 150, i.e., 100 samples. Random resampling is performed from the synthetic

database to obtain the Bootstrap sample for each class. Continuing the example, for class "0", 50 samples will be randomly selected from the synthetic database for class "0", and for class "1", 100 samples will be randomly selected from the synthetic database. The Bootstrap samples for each class are concatenated to form the complete Bootstrap sample. In our example, all 50 samples from class "0" and 100 samples from class "1" are concatenated to form a single Bootstrap sample. This sample contains a combination of samples from both classes, keeping the original class proportions. Bootstrap sample generation occurs for each iteration of the process, i.e., each time a new set of synthetic data is desired to be generated.

4.2.4 Blending synthetic data

After generating each Bootstrap sample, the data from each sample is randomly shuffled to ensure variability and remove any order correlations. The algorithm generates a list of data frames representing the synthetic Bootstrap samples, maintaining the class distribution and initial features.

4.3. Postprocessing

The post-processing techniques adopted by the “Fusionstrap” method consist in the elimination of extreme values using both the z-score and the IQR (Interquartile Range) method, from the synthetic data sets generated by means of the hybrid fusion between Bootstrap methods and Gaussian Copula algorithm. These techniques were implemented to ensure the integrity and representativeness of the synthetic data, before applying additional transformations, such as Bootstrap Rotation, to improve the diversity and accuracy of the resulting datasets.

4.3.1 Eliminating extreme values (outliers) from the synthetic data set

Outliers are those values that are much different from the mean or the other values in the data set and can adversely affect subsequent analyzes and models. To identify and remove these values, the framework uses the z-score statistic and the Interquartile Range (IQR) method [100]. The algorithm of this process assumes [100]:

- Going through each synthetic data set: For each data set in the synthetic_datasets list created in the previous steps, only the numeric columns are selected for analysis.
- Identifying and removing outliers: The method constructs the feature matrix X (independent variables) and uses the Python function np.percentile to calculate the 25th and 75th percentiles (Q1 and Q3) for each column of the dataset's feature matrix. This is done by specifying the axis=0 argument to the np.percentile function. When axis=0, the np.percentile function calculates the percentiles for each column separately, thus allowing the distribution of each variable to be evaluated. Thus, in the result, we have Q1 and Q3 for each column of X, which allows us to calculate the IQR for each variable. The threshold for identifying extreme values is set to 1.5, but can be adjusted according to the needs and specifics of the data set. IQR is then calculated as the difference between Q3 and Q1. To identify the extreme values in the data matrix: by applying the IQR threshold and the Q1 and Q3 values, it is determined which values are considered extreme by comparing them with the defined threshold. The formula used is:

  Outliers formula = (X < (Q1 - threshold * IQR)) | (X > (Q3 + threshold * IQR))

  By using the comparison operators (< and >), it is checked whether each value in the dataset array is less than Q1 minus threshold * IQR or greater than Q3 plus threshold * IQR.
- The result is an "outliers" matrix of the same form as the dataset, where values that exceed the thresholds are marked as "True".
- Filtering and removing outliers: The “out-liers matrix” is used to filter the outliers in the data set. Finally, the dataset without extreme values (filtered dataset) is added to a “no_outlier_synthetic_datasets” list.

4.3.2 Combining via Bootstrap Rotation

The bootstrap algorithm with rotation is a technique used in statistics for generating synthetic datasets with the aim of improving the diversity and robustness of estimates by rotating observations in datasets [101]. “Fusionstrap” uses an adaptation of Bootstrap Rotation to combine the 150 synthetic samples filtered in the previous step into a single dataset of the same size as the original dataset. The method involves combining observations from filtered data sets in a round-robin fashion so that each observation has an equal chance of being included in the sample.

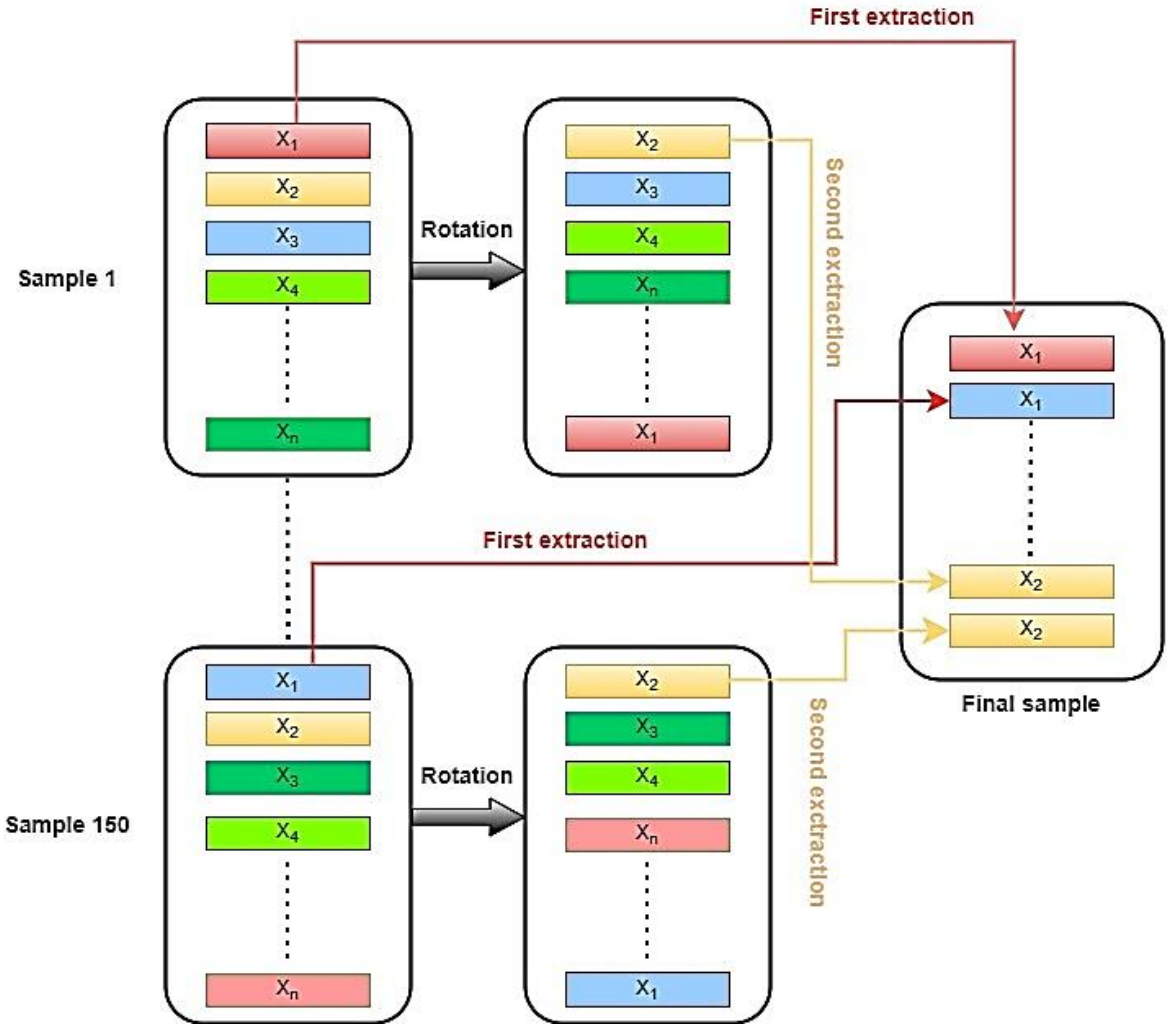
The “Fusionstrap” algorithm for combining by rotation the observations from the 150 samples involves the following steps (Figure 18):

- A final sample is initialized as empty.
- Each iteration consists of:
  - One observation is extracted from each sample and added to the final sample.
  - Circularly rotate the observations from each sample. This means that the observations that were extracted are put back in each sample of provenance in the last position.
- The iterative process continues until the final sample reaches the size of the original data set.

This approach can help create a more varied and balanced sample in terms of data distribution.

Figure 18

Combining with Bootstrap Rotation



Note. The figure shows the process of the first iteration, the rotation of the observations in each sample and the first two extractions. The observations from each sample were marked with X1...., Xn.

Chapter 5
Experiments

This chapter aims to expose various experimental configurations applied in this research. All experiments will be run using the datasets presented in Section 5.1. Before detailing the experiments, we give a brief overview of the context of the data. Next, we will focus on the actual experiments. The fundamental purpose of the experimental setups is to evaluate whether the methods of the proposed framework can lead to an improvement in the quality of the generated synthetic data. In addition to this aspect, it is aimed to investigate the framework's ability to improve the fairness of the results by addressing the problem of unbalanced data in the dataset. In the synthetic data evaluation process, we will use the methods and metrics analyzed in detail in Chapter 3. This approach ensures a robust and comprehensive analysis of the data quality generated by the proposed framework. More details about the experimental results can be found in **Appendix A.**

5.1 Datasets

The rigors of our research were guided by a careful selection of datasets to use for experiments. We opted for three distinct datasets, each with significant characteristics relevant to our research objective. The choice of these data sets was influenced by the following considerations, which emphasize the importance and concordance with the criteria and objectives of the present research:

- **US Census:** The choice of this data set was guided by its comprehensiveness and diversity, reflecting the different demographic, social and economic aspects of the population of the United States of America. Because of its size and complexity, this dataset allows us to assess how well our framework can address the diversity and complexity of information to generate relevant and realistic synthetic data.
- **Diabetes Prediction**: The Diabetes Prediction dataset was selected due to its medical nature and the critical importance of accuracy in medical analyses. By applying our framework to this dataset, we aim to demonstrate its ability to generate synthetic data that preserves essential characteristics of the study population so that medical analyzes remain robust and valid.
- **AIDS:** AIDS: We selected this data set given its complex and sensitive nature, as well as the stigmatizing connotations associated with HIV/AIDS. By applying our method to this data set, we aim to highlight our framework's ability to handle sensitive data while protecting individual privacy while providing relevant information for health-related analyses. Another reason for choosing this data set is to allow a direct comparison of the results obtained with the results of a pre-existing study on the same data collection, which used the AVATAR method [94]. This approach helps to validate and evaluate the relative performance of our method compared to other existing approaches.

Another valuable aspect that we have considered in choosing the data sets is their varied size. The datasets range in size from large to small, reflecting the diversity of sizes that our framework might encounter in a wide range of research contexts. This approach allows us to examine the behavior and performance of the framework as a function of dataset size, thereby contributing to a deeper understanding of how the generated synthetic data can be influenced by the variability of input data sizes. Moreover, the selected datasets allow testing the effectiveness of our proposed framework in addressing class imbalance issues, providing the opportunity to assess to what extent the generated synthetic data can contribute to improving equity and performance in the context of datasets characterized by unequal class distributions.

5.1.1 US Census

These data were collected from the 1994 Census Bureau database with the coordinated efforts of Ronny Kohavi and Barry Becker [102]. The initial set contains a total of 48,842 rows of data. The data set extraction process followed carefully established criteria, ensuring a consistent and valid set of records.

These criteria included the following conditions: age of individuals to be greater than 16 years, adjusted gross income to exceed $100, assigned final weight (AFNLWGT) to be greater than 1, and number of hours worked per week to be greater than 0 [102]. The main objective of this dataset is to build models with the ability to predict if a person's annual income exceeds $50,000.
The set consists of a target variable “income” and 14 predictor variables, which are a mixture of categorical, ordinal, and numeric data types representing various socio-economic and demographic characteristics of the individual. A description of the variables can be found in Appendix A. All these variables are used in predicting the target variable “income,” indicating whether an individual earns more than $50,000 annually.
This data set is of moderate size and shows significant class imbalance, as the number of records associated with people earning more than $50,000 per year is lower than the number of records associated with people earning less than or at most $50,000 per year year. year. Thus, there are two class values ">50K" and "<=50K", the classes being unbalanced, with a trend towards the class label "<=50K". This can influence the performance evaluation of prediction models and is an ideal scenario for testing the framework's ability to address the class imbalance problem by generating synthetic data. Also, the set contains missing values marked with "?". The experiments in this research will be performed on a number of 32,561 recordings extracted from this set.

Table 5
US census Database

Instances	32.561
Attributes	14

5.1.2 Diabetes Prediction

This dataset contains 4000 records from the US National Institute of Diabetes and Digestive and Kidney Diseases database, the primary data source being electronic health records (EHR). EHRs are digital versions of patients' health records that contain information about their medical history, diagnosis, treatment and outcomes. EHR data is collected and stored through surveys, medical records, and laboratory tests by health care providers, such as hospitals and clinics, as part of their routine clinical practice. The objective is to predict, based on diagnostic measurements, whether a patient has diabetes [104].

Table 6
Diabetes Prediction Database

Instances	4.000
Attributes	9

The set contains the target variable "diabetes" which, in the context of predicting diabetes, indicates whether the person has diabetes (1) or not (0) and 8 other predictor variables, which are a mixture of categorical and numerical data types. The nine variables of the data set are described in **Appendix A**. The data set used for the experiments has a small size and there is the possibility of presenting class imbalances. This will be analyzed through the experiments carried out in this research.

5.1.3 AIDS

This HIV infection dataset contains information on 2139 patients and comprises 26 variables. These individuals participated in a clinical trial documented in a 1996 publication in the New England Journal of Medicine. Conducted by Hammer and collaborators, the study comprised four distinct treatment cohorts. The primary aim of the investigation was to evaluate patient survival and to observe any potential 50%

reduction in CD4+ cell count [105]. The data set is small in size, contains only numerical variables **(Appendix A)** and has no missing values.

Table 7
AIDS Database

Instances	2.139
Attributes	26

5.2 Experimental setup

5.2.1 Preprocessing

This first step of the experiment aims to ensure that the preprocessing function within "Fusionstrap" is working properly, that the data is clean, and that any subsequent changes in performance can be validly attributed to the developed method.

5.2.1.1 Replacement of missing values

To obtain assurance about the quality and validity of the data we will use later, we will use the "US Census" data set. This dataset was selected due to the presence of missing values in multiple variables, offering an opportunity to assess the efficacy of the preprocessing function in addressing such scenarios.
The preprocessing process will involve replacing missing values with the methods specified in section 4.1 of the method. In the "US Census" set, we identified a total of 4266 missing values distributed in different variables, as follows:

Workclass: 1836 missing values
Education.num: 2 missing values
Occupation: 1843 missing values
Capital Loss: 1 missing value
Hours per Week: 1 missing value
Native Country: 583 missing values

To manage the missing values, the average of the existing values in the numerical variables (Education.num, Capital Loss, Hours per Week) was used, and, in the case of the categorical/non-numerical variables (Workclass, Occupation, Native Country), the missing values were replaced with the mode (the value that occurs most often). Using the mean helps maintain data consistency and avoid skewing the distribution, and mode choice helps preserve dominant features and minimize the impact on the data distribution.
In the "AIDS" set, we identified a total of 797 missing values distributed in the numerical variable "cd496".

5.2.1.2 Clean data evaluation

The evaluation of this preprocessing process will consist of the following steps:

- Basic Statistics Check: We will compare the basic statistics of the original data set with those of the cleaned data set. This will allow us to see if the distribution and values have remained within reasonable limits.
- Correlation Analysis: We will examine the correlations between variables in the cleaned data set. This analysis will help us identify trends, significant relationships, or possible anomalies in the data.
- The evaluation results will be presented in the form of comparative tables and graphical visualizations such as heat maps. These will provide a visual and easy-to-understand insight into the differences and correlations in the raw and processed data.

5.2.1.3 Definition of layers

This part of the experiment aims to assess the distribution and imbalance of classes in datasets, providing a basis for further generation of synthetic data to correct imbalance and improve data quality. The method approached by the Fusion strap framework in this regard is the one described in Section 4.2. All three data sets (US Census, Diabetes Prediction and AIDS) will be subjected to the experiment, and the results (class distribution and imbalance percentage) will be presented in the form of tables and histograms for each variable. Variables with high percentages of imbalance indicate significant inequity between classes.

5.2.2 Generating synthetic data: "Fusionstrap" vs other methods

The central part of the experiment in this research focuses on an exhaustive comparison between the "Fusionstrap" framework and other synthetic data generation methods. The data synthesis techniques used for this comparison, namely CTGAN and SYNTHPOP, were detailed in Section 3.3 of the study. In this experiment, we aim to generate synthetic data both through the "Fusionstrap" framework and using the alternative methods mentioned above. This endeavor will be carried out across the full range of data sets used in this study, i.e., US Census, Diabetes Prediction and AIDS. The evaluation of the quality of the data generated by the three approaches ("Fusionstrap", CTGAN and SYNTHPOP) will be carried out through a thorough comparison, based on the results obtained following the application of the evaluation methods and metrics described in detail in Sections 3.4 and 3.5 of the paper. This experiment is a focal point in analyzing the performance of the "Fusionstrap" method in the context of other synthetic data generation approaches.

5.2.3 Postprocessing

Postprocessing will involve removing outliers from the 150 samples generated with "Fusionstrap" and then applying the bootstrap rotation to the cleaned samples, resulting in a single synthetic dataset of the size of the original dataset. This phase aims to ensure that the synthetic data aligns with real-world patterns and distributions. The methods were detailed in Sections 4.3.1 and 4.3.2, and the results will be presented in the next chapter.

Chapter 6
Results

This chapter brings to the fore the results obtained from the experiments carried out as part of our research. The main objective of these experiments was to assess the effectiveness of the "Fusionstrap" framework in generating synthetic data and addressing class imbalance across three distinct datasets: US Census, Diabetes Prediction, and AIDS. Additionally, the merits and drawbacks of the proposed "Fusionstrap" framework will be deliberated. In this thesis, summaries of the results were presented. More details can be found in **the Appendix.** We will detail the results for each individual experiment, providing context and the appropriate interpretation of the data obtained.

6.1 Evaluation of the Preprocessing Function

In this experiment, we focused on evaluating the preprocessing function of the “Fusionstrap” framework. We chose the US Census data set, which contained missing values for several variables, to highlight how the “Fusionstrap” framework addresses this issue. To replace missing values from the US Census data set we used the methods described in section 4.1.

6.1.1 Keeping basic statistics in the cleaned dataset

We compared the baseline statistics of the original data set with those of the cleaned set to assess the validity of the preprocessed data. In the case of the original US Census set, 4262 values are missing for categorical variables (“workclass”, “occupation”, native.country”) and 4 values for numeric variables (“education.number”, “capital.loss”, “hours.per.week”).
According to the data in Table 8, which shows the results regarding the differences between the statistics of the cleaned and the original set for the numerical variables, the following conclusions can be formulated:

- The number of records without missing values from the data set (count) increased by 2 records for the variable "education.num", by 1 record for "capital.loss" and by 1 record for "hours.per.week". This confirms that all missing values from the original set have been substituted for these variables. Therefore, the cause of these differences is the input of new values in place of missing data.
- The difference of -0.00267866 in the case of the "capital.loss" variable indicates a fairly small decrease of the average in the cleaned data set compared to the initial one. This difference can be attributed to the way missing values were replaced in this data set and can be considered a normal fluctuation in the context of the data used for this experiment.
- In terms of standard deviation (std) very small differences such as -0.000079013 and -0.000189607 can be considered as normal fluctuations in the data. Very small differences in the standard deviation of the variables between the original and the cleaned data set may be caused by the process of replacing missing values by the preprocessing method. Replacing missing values can introduce small variations in the data, depending on the values they are replaced with. In addition, the replacement process may affect the standard deviation of the data negligibly, especially in cases where missing values are rare and/or uniformly distributed. Also, there could be minor differences caused by calculation errors or mathematical approximations in preprocessing time. If the original and cleaned data are essentially very similar, then any difference in standard deviation could be the result of small rounding or data manipulation errors. In conclusion, although there are numerical differences, they are negligible and should not have a significant impact on the interpretation of the data.
- The "min" statistics (the smallest value of each variable), the 25th percentile (the value below which 25% of the data is located), the 50th (the value below which 25% of the data is located) and the 75th percentile (the value below which 75% is located from the data), as well as "max" (the highest value of each variable) were not affected by preprocessing.

Table 8: *Differences between the numerical statistics of the cleaned set and the original set*

	age	fnlwgt	education.num	capital.gain	capital.loss	hours.per.week
Count difference	**0**	**0**	**2**	**0**	**1**	**1**
mean original	38.58	189778	10	1078	87.21977887	40.44
mean clean	38.58	189778	10	1078	87.21710021	40.44
mean difference	**0**	**0**	**0**	**0**	**-0.00267866**	**0**
std original	13.64	105549.98	2.572608256	7385.29	402.6808776	12.34709879
Std clean	13.64	105549.98	2.572529243	7385.29	402.6749840	12.34690918
Std difference	**0**	**0**	**-0.000079013**	**0**	**-0.0058936**	**-0.000189607**
min original	17	12285	1	0	0	1
min clean	17	12285	1	0	0	1
min difference	**0**	**0**	**0**	**0**	**0**	**0**
25% original	28	117827	9	0	0	40
25% clean	28	117827	9	0	0	40
25% difference	**0**	**0**	**0**	**0**	**0**	**0**
50% original	37	178356	10	0	0	40
50% clean	37	178356	10	0	0	40
50% difference	**0**	**0**	**0**	**0**	**0**	**0**
75% original	48	237051	12	0	0	45
75% clean	48	237051	12	0	0	45
75% difference	**0**	**0**	**0**	**0**	**0**	**0**
max original	90	1484705	16	0	0	99
max clean	90	1484705	16	0	0	99
max difference	**0**	**0**	**0**	**0**	**0**	**0**

The analysis of the results in **Table 9**, which shows the differences between the statistics of the cleaned set and the original set for the categorical variables, highlights the following aspects:

- "count": the results indicate differences in the number of instances between the cleaned set and the original set for the variables "workclass" (1836 values), "occupation" (1843 values) and "native.country" (583 values). These differences confirm the complete replacement of missing values from the original set.
- a "False" result (eg. for the "top" variable) or a zero result means that the number of instances is the same in both sets for the analyzed variables.
- "unique": there are no differences between the number of unique values between the original set and the cleaned set (the difference result is zero for all variables).
- "frequency": after replacing the missing values with the mode, the frequencies of the most common work class (Private), the most common occupation (Prof-specialty) and the most common country of origin (USA) increased compared to the original set with 1836, 1843, respectively 583 courts. This is due to the replacement of missing values with fashion, which led to an increase in these frequencies.

Table 9: *Differences between the categorical statistics of the cleaned set and the original set*

	workclass	education	marital status	occupation	relationship	race	sex	native country	income
count orig	30725	32561	32561	30718	32561	32561	32561	31978	32561
count clean	32561	32561	32561	32561	32561	32561	32561	32561	32561
count dif	1836	0	0	1843	0	0	0	583	0
unique orig	8	16	7	14	6	5	2	41	2
unique clean	8	16	7	14	6	5	2	41	2
unique dif	0	0	0	0	0	0	0	0	0
top orig	Private	HS-grad	Married	Prof-speciality	Husband	White	Male	US	<=50k
top clean	Private	HS-grad	Married	Prof-speciality	Husband	White	Male	US	<=50k
top dif	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
freq orig	22696	10501	14976	4140	13193	27816	21790	29170	24720
freq clean	24532	10501	14976	5983	13193	27816	21790	29753	24720
freq dif	1836	0	0	1843	0	0	0	583	0

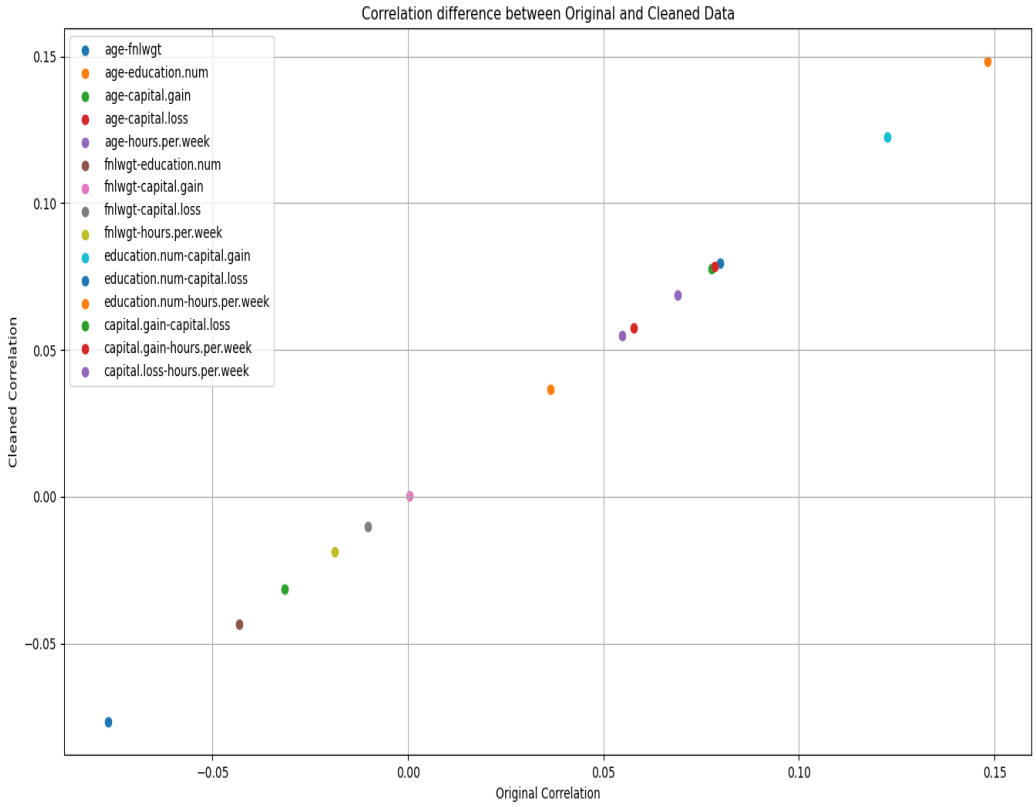
Overall, the “Fusionstrap” method demonstrated efficiency and consistency in handling missing values through the preprocessing method, while preserving the overall integrity of the data and minimizing significant influences on the distribution or underlying characteristics. These findings suggest that preprocessing using the “Fusionstrap” method can be an effective approach to prepare data for later stages of research.

6.1.2 Keeping correlations in the clean dataset

In our analysis process on the preservation of correlations in the cleaned data set compared to the original one, we adopted the method detailed in Section 4.1. We started by determining the correlations between the variables of the initial set. Utilizing a dedicated Python code, we conducted computations for the Pearson correlation coefficient for numerical variables and the Cramer’s V correlation coefficient for nominal/mixed variables. For nominal values, we created contingency tables between these variables and a reference variable, then calculated chi-square values and applied the formula for Cramer's V correlation coefficient. We repeated the calculations of the Pearson correlation coefficient for the numerical variables and of the Cramer’s V correlation coefficient for the nominal/mixed variables also in the case of the cleaned dataset. In the next step, we calculated the absolute difference between the values of the Pearson and Cramer s V coefficients corresponding to the cleaned data set and the values of the original set. Based on these differences we constructed a correlation matrix, highlighting the changes in the correlations of the variables. The correlation matrix was generated using bias correction. For a visual understanding, we created scatterplots and radar plots (Figures 19 and 20), illustrating the changes to the correlations following the data cleaning process.

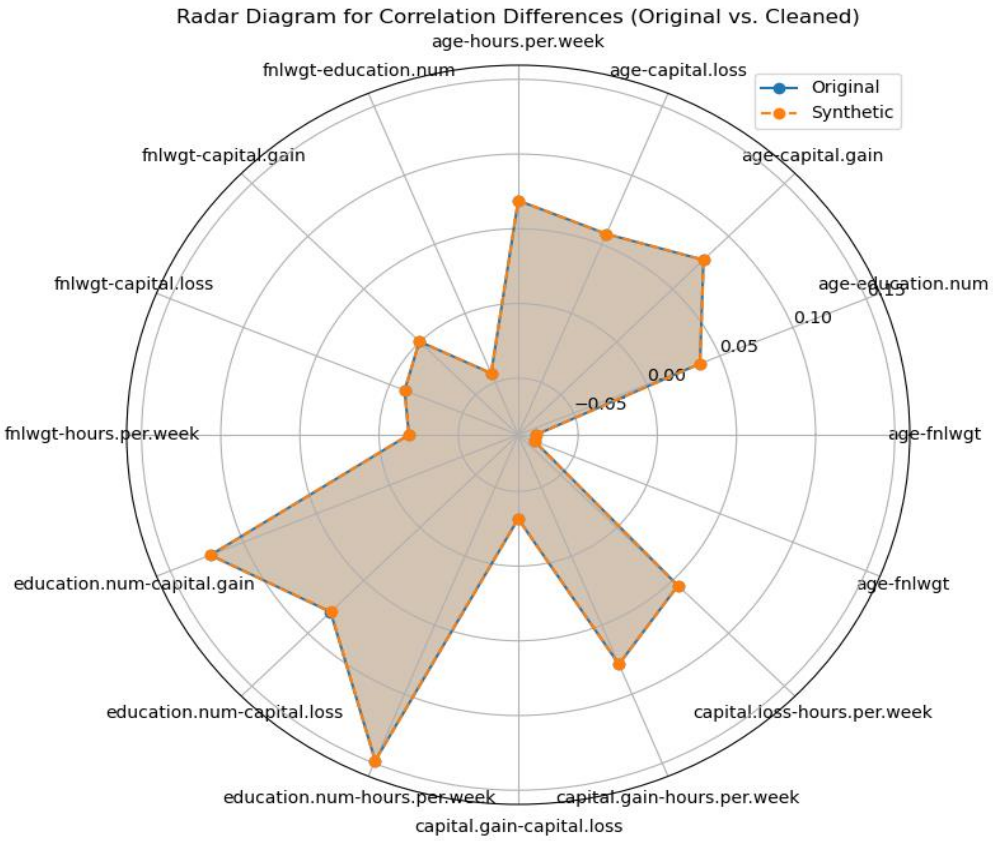
Figure 19

Correlation scatterplot diagram



The scatterplot diagram indicates a uniform distribution of points around the diagonal line that starts increasing from the intersection of the axes, and the values of the coefficients (-0.05:0.15) are close to 0. This suggests that the changes between the correlations are minimal or insignificant and that there is a close correspondence between the cleaned data and the original data in terms of correlations. We can conclude, therefore, that the data cleaning process did not have a major impact on the structure of the correlations between the variables.

Figure 20
Correlation radar diagram



The radar diagram indicates a perfect overlap of the polygons formed around the center by connecting the points that mark the values for each correlation coefficient. This indicated, once again, that the changes introduced into the cleaned data set were minimal in terms of correlations and that the structure of the relationships between the variables remained almost unchanged.
After careful analysis of the results of the data preprocessing experiment, we can conclude that the cleaning function implemented in “Fusionstrap” was able to maintain the integrity of the original data, both from the point of view of basic statistics and correlations. This effectiveness in preserving the essential features of the data gives us confidence in using the cleaned data set in further analyzes without significantly distorting the results or interpretations.

6.2 Definition of layers

The present experiment focuses on assessing the distribution and inequity of the classes in the datasets, with the goal of underpinning the subsequent generation of synthetic data to correct these imbalances and improve data integrity. To achieve this goal, the “Fusionstrap” framework adopts the method detailed in Section 4.2 of this study. The results of the experiment which include the percentage of imbalance are presented in the form of tables and histograms for each individual variable (Figure 21, Figure 22 and Figure 23).

Figure 21

Percentage of imbalances in the US census set

VARIABILE	ORIGINAL
workclass	99.97
education	99.51
marital.status	99.85
occupation	99.85
relationship	92.56
race	99.03
sex	50.57
native.country	100
income	68.28

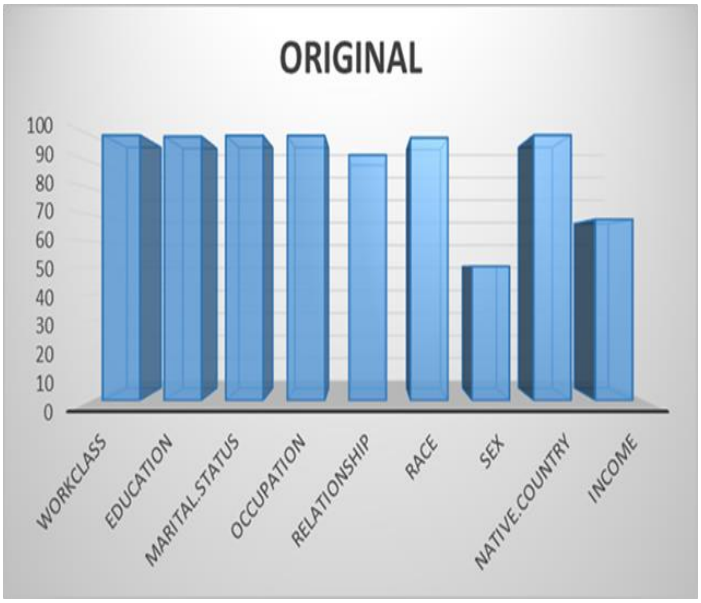


Figure 22

Percentage of imbalances in the Diabet Prediction set

VARIABILE	ORIGINAL
gender	99.96
hypertension	92.04
heart_disease	95.75
smoking_history	88.5
diabets	91.45

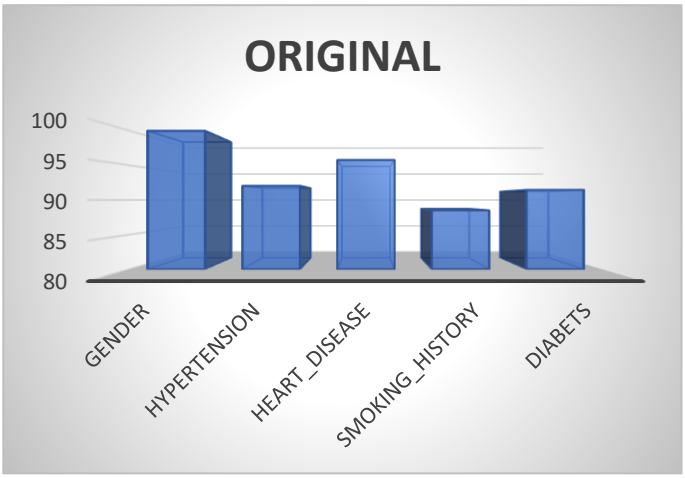
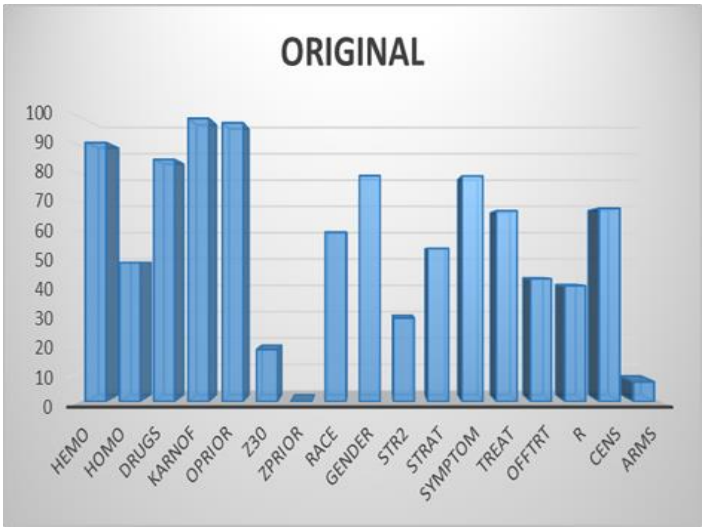


Figure 23
Percentage of imbalances in the AIDS set

VARIABILE	ORIGINAL
hemo	90.81
homo	48.73
drugs	84.88
karnof	99.29
oprior	97.75
z30	18.27
zprior	0
race	59.46
gender	79.22
str2	29.29
strat	53.72
symptom	79.08
treat	66.89
offtrt	43.07
r	40.61
cens	67.8
arms	6.95



The results presented in visual and numerical forms (Figure 21, Figure 22 and Figure 23) constitute a valuable resource for understanding the gradient of inequality within each data set and will influence subsequent data manipulation decisions to obtain results more accurate and fairer. Variables with significant percentages of inequity indicate significant discrepancies between classes, thus suggesting that they require greater attention in the process of generating synthetic data. For example, the table in Figure 21 shows the degree of class imbalance for each variable in the US census data set. Class imbalance refers to the significant difference between the frequency of different classes of a variable. Below is the interpretation of these percentages, highlighting the specific nature of the imbalance for each variable:

- the percentage of 99.97% for "workclass" signals a pronounced disproportion in the distribution of occupational classes. This assumes that certain categories of work are much more prevalent than others;
- the percentage of 99.51% for "education", similar to "workclass", indicates a significant imbalance in the distribution of education levels, some categories being much more frequent;
- the percentage of 99.85% for "marital.status" highlights a significant disproportion in the distribution of marital statuses, suggesting a higher prevalence of some marital statuses compared to others;
- the percentage of 99.85% for "occupation", similar to "workclass" and "education", indicates a notable imbalance in the distribution of occupations, where certain types of occupations predominate;
- notable differences between the categories are also presented by the variables: "relationship" (92.56%) and "race" (99.03%);
- the "sex" variable registers a percentage of 50.57% that approaches 50%, which indicates a relatively fair distribution between the sexes, with a more uniform approach to the classes (men and women).
- for "native.country" (100%), all data belong to the same native state, signaling a lack of diversity in this dimension of the dataset.

- the percentage of 68.28% for "income" indicates a relatively balanced distribution between income categories, with a slight inequality, indicating a possible prevalence of one income category.

These percentages provide essential information about the distribution of classes in each variable, which is crucial for assessing class imbalance and identifying areas where intervention is needed to ensure a more equitable distribution in the data set. How these class imbalances will be handled will be assessed in the next experiment.
In addition to generating synthetic data, this experiment also aims to investigate and compare the effectiveness of the “Fusionstrap” method in addressing class imbalances with other methods or techniques for managing these inequities. By applying different imbalance management strategies and comparing the results to the original data sets, we will be able to determine whether the “Fusionstrap” method makes significant improvements in correcting these discrepancies and obtaining more balanced data.

6.3 Generating synthetic data: “Fusionstrap” vs other methods

The central experiment of this research aims at an exhaustive evaluation of the “Fusionstrap” method compared to other synthetic data generation techniques. This experiment aims to compare the results obtained using “Fusionstrap” with those generated by alternative methods such as CTGAN and SYNTHPOP. Synthetic data quality assessment will be performed on the datasets chosen for this study: US Census, Diabetes Prediction, and AIDS. By applying the evaluation methods and metrics described in Sections 3.4 and 3.5 of the paper, we will analyze the performance and effectiveness of each approach in generating synthetic data. For this experiment, we tested unbalanced data sets to see how the “Fusionstrap” method could handle class imbalance problems compared to other methods designed for this purpose.

6.3.1 Evaluation of utility

Applying the evaluation methods and metrics described in Section 3.4, we will analyze the performance and effectiveness of each approach to generate coherent and viable synthetic data (utility). The identification of the method that produces the best results will be achieved through the comparative analysis of the results. In addition to the aforementioned analyses, to comprehensively assess the usefulness of the synthetic data, we will perform an additional evaluation of two key statistics of interest from the AIDS dataset. These statistics are the survival curve and the hazard ratio. This supplementary evaluation step is designed to confirm the suitability and coherence of the synthetic data produced by the "Fusionstrap" method concerning the inherent characteristics of a sensitive dataset.

6.3.1.1 Hellinger Evaluator

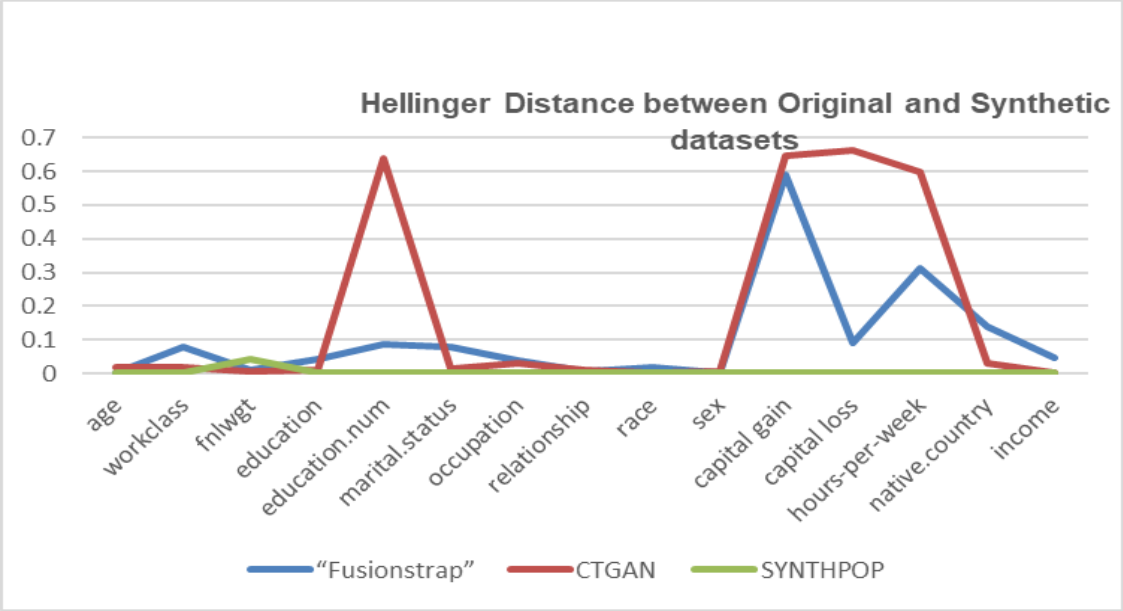
Figure 24 shows the graphical representation of the Hellinger distance values (**Appendix B**) obtained by the three data synthesis approaches from the US Census data set.

The Hellinger distance is utilized to gauge the resemblance between the probability distributions of synthetic and real data, ranging from 0 to 1 (Section 3.4). A value close to 0 indicates a strong fidelity of synthetic data in mimicking the actual distribution, signifying a significant utility in the synthesis process. Conversely, a value near 1 suggests notable disparities between the probability distributions of synthetic and real data, implying a reduced effectiveness of synthetic data in analytical or modeling scenarios.
The comparative analysis of Hellinger distances (Figure 24) indicates the following key points regarding the performance of "Fusionstrap" compared to CTGAN and Synthpop:

- "Fusionstrap" stands out by small distances for critical variables such as "age","workclass""education.num", "occupation", "race" and "sex". This indicates a significant similarity between the synthetic and real distributions for these essential features.

- “Fusionstrap” records larger distances for variables such as capital gain, capital loss, and hours-per-week. This indicates a lower performance of "Fusionstrap" in ensuring the utility of synthetic data for these features. However, Fusionstrap obtains better values for these variables than CTGAN.
- Despite some solid results, "Fusionstrap" is outperformed by Synthpop in terms of overall similarity between synthetic and real distributions. CTGAN ranks third, indicating a lower performance compared to the other two methods.

Figure 24
Hellinger distance for the US census synthetic set

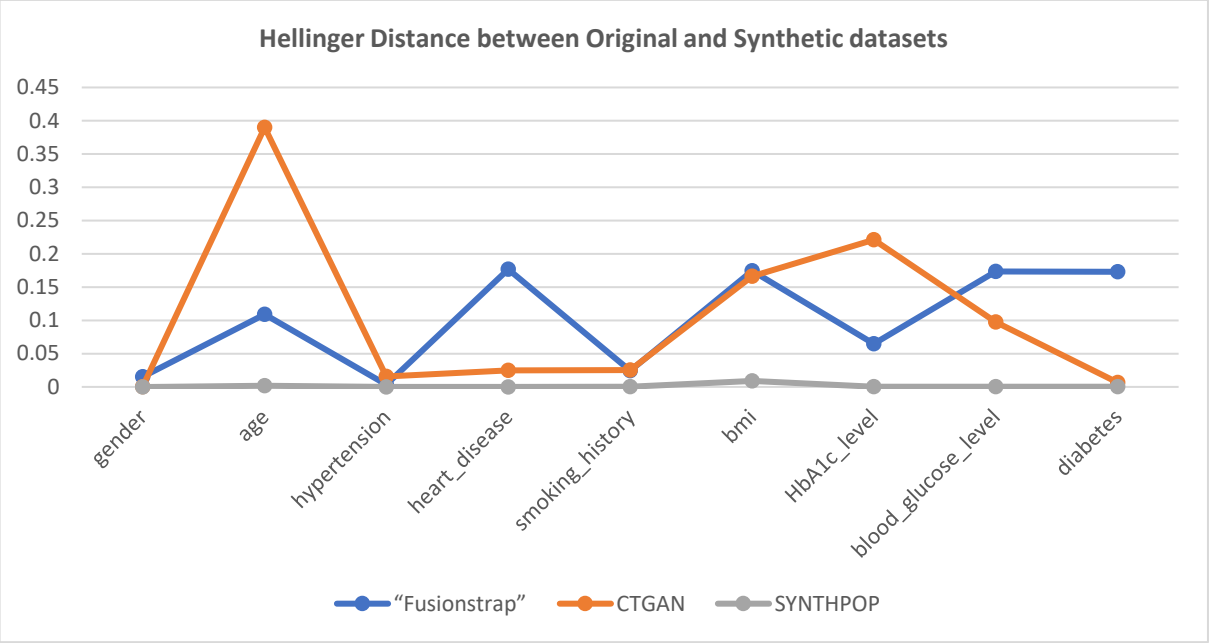


Analysis of Hellinger distances for the Diabetes Prediction data set **(Appendix B)** reveals significant differences in performance between "Fusionstrap", CTGAN and Synthpop (Figure 25). In evaluating this comparison, the following relevant aspects can be highlighted:

- “Fusionstrap” shows significant efficiency in capturing the distributions for “gender”, “hypertension”, “smoking_history” and “HbA1c_level” variables, recording significantly smaller distances than the CTGAN method.
- A point of vulnerability of "Fusionstrap" is illustrated by the significantly higher distances for the variables "heart_disease", "bmi", "blood_glucose_level" and "diabetes" compared to the "Synthpop" method, signaling a lower similarity in the distribution of these key features.
- This analysis reveals that, despite notable performances of "Fusionstrap" for some variables, the method fails to outperform Synthpop, which remains predominant in ensuring similarity of synthetic distributions to real ones in the specific context of the Diabetes Prediction dataset. CTGAN ranks last, highlighting a lower relative effectiveness for this particular data set.

Figure 25

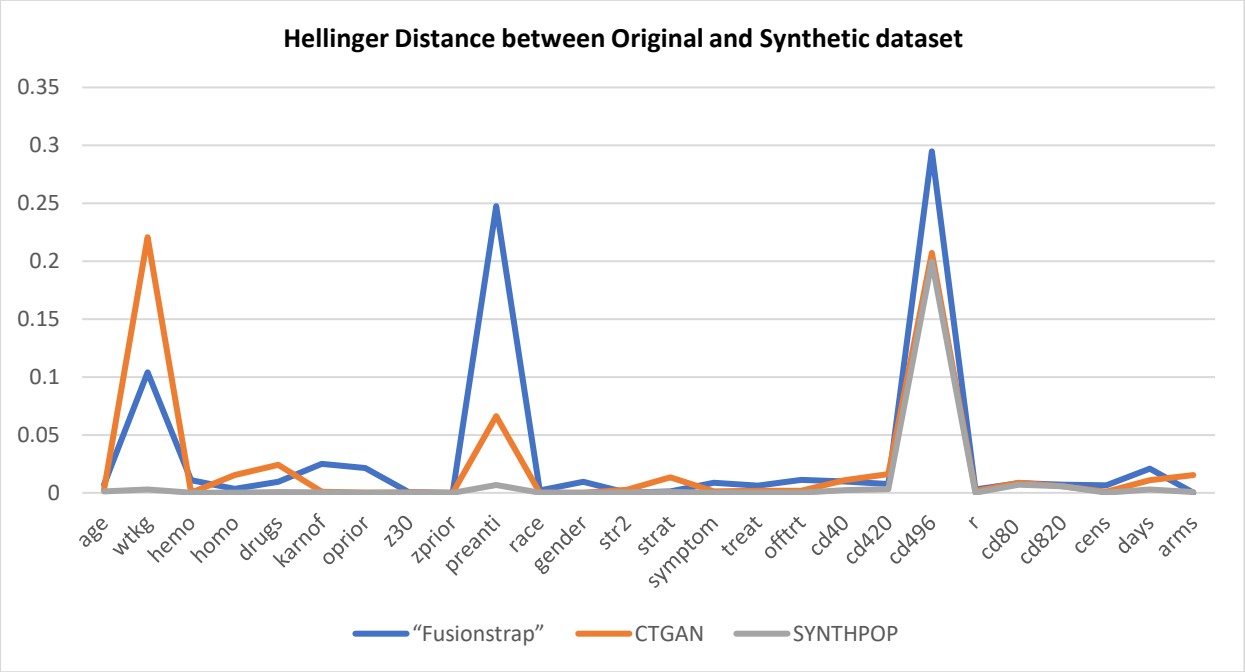
Hellinger distance for the Diabetes Prediction synthetic set



From Figure 26, representing the Hellinger distances for the SIDA set **(Appendix B)**, the following essential aspects emerge:

- "Fusionstrap" exhibits exceptional efficacy in preserving the similarity of distributions across various variables, showcasing minimal or even negligible distances, particularly in "gender," "zprior," and "arms."
- For the "preanti" and "cd496" variables, "Fusionstrap" scores significantly higher distances than Synthpop and CTGAN, indicating relatively poor performance for these features.
- This analysis indicates a performance of "Fusionstrap" comparable to that of Synthpop performance for most variables. However, certain variables are also identified where "Fusionstrap" outperforms Synthpop. "CTGAN" remains in last place in most cases, indicating a lower performance in generating synthetic data for this data set.

Figure 26
Hellinger distance for the AIDS synthetic set



The overall conclusion of the analysis reveals that "Fusionstrap" stands out for its efficiency in capturing the distributions for many key variables on the three selected datasets. However, weak points were also identified, especially compared to Synthpop, which continues to be the leader in terms of overall similarity between synthetic and real distributions. CTGAN ranks lower in the overall performance ranking. The selection among these approaches should consider the specific characteristics of the dataset, the variables of concern, and the analysis objectives. It is crucial to weigh both the advantages and disadvantages of each method, as there is no universally superior approach for all variables.

6.3.1.2 Correlation Evaluator

The methods and metrics used in this experiment to evaluate and compare the performances of the three approaches ("Fusionstrap", CTGAN and SYNTHPOP) in terms of preserving the correlations between variables with respect to the original set were presented in detail in Section 3.4 of the paper. Depending on the dimensionality of each dataset, the outcomes were depicted using radar diagrams or scatterplot diagrams for optimal visual representation. Additionally, to comparatively assess the utility of synthetic data, the univariate distributions were examined for the three datasets across the three approaches ("Fusionstrap," CTGAN, and SYNTHPOP) (Appendix C). Through this comparison of the univariate distributions between synthetic and original data, we can discern the degree to which synthetic data accurately mirrors the distribution patterns observed in the authentic dataset.

- ***US census***

Figure 27 presents radar charts for the U.S. Census dataset, illustrating the differences between correlation coefficients of synthetic data generated with "Fusionstrap" (a), CTGAN (b), and SYNTHPOP (c) compared to the coefficients of the original data. The values underlying these charts are provided in **Appendix B**. The differences between synthetic-original correlation coefficients reflect how well the synthetic models manage to reproduce relationships between variables in the original dataset. In interpreting these differences, the following aspects can be observed:

- Value range **(Appendix B)**:

Minimum: "Fusionstrap": 0.005309023, CTGAN: 0.000537373, SYNTHPOP: 0.000037300
Maximum: "Fusionstrap": 0.634494573, CTGAN: 0.960033808, SYNTHPOP: 0.039659307

- Percentage of cases with minimum values **(Appendix B)**:

Fusionstrap: 7%
CTGAN: 10%
SYNTHPOP: 87%

- Interpretation on variable groups **(Appendix B)**:

Compared to SYNTHPOP and CTGAN, “Fusionstrap” seems to fall within a medium-performance range. SYNTHPOP records the smallest differences in most variables, indicating a better ability to accurately reproduce relationships between variables. However, “Fusionstrap” has some significant advantages in certain scenarios, such as variable pairs “capital.loss+hours.per.week,” “education.num+hours.per.week,” “hours.per.week+race,” “marital.status+race,” “native.country+occupation,” and “sex+workclass,” where it records significantly smaller differences than CTGAN and performs comparably to SYNTHPOP. The significantly large differences for pairs “age+relationship,” “education+education.num,” and “relationship+sex” indicate a potential weakness in “Fusionstrap” in adequately reproducing these relationships. These examples illustrate an area where the model can be improved to achieve SYNTHPOP-like performance.

- General conclusion:

SYNTHPOP stands out by recording the smallest differences in 87% of variables, indicating better performance in accurately reproducing relationships between variables. Fusionstrap falls within a medium range, with the smallest differences in 7% of variables. CTGAN, despite recording relatively larger differences, demonstrates better performance than Fusionstrap in 10% of variables.

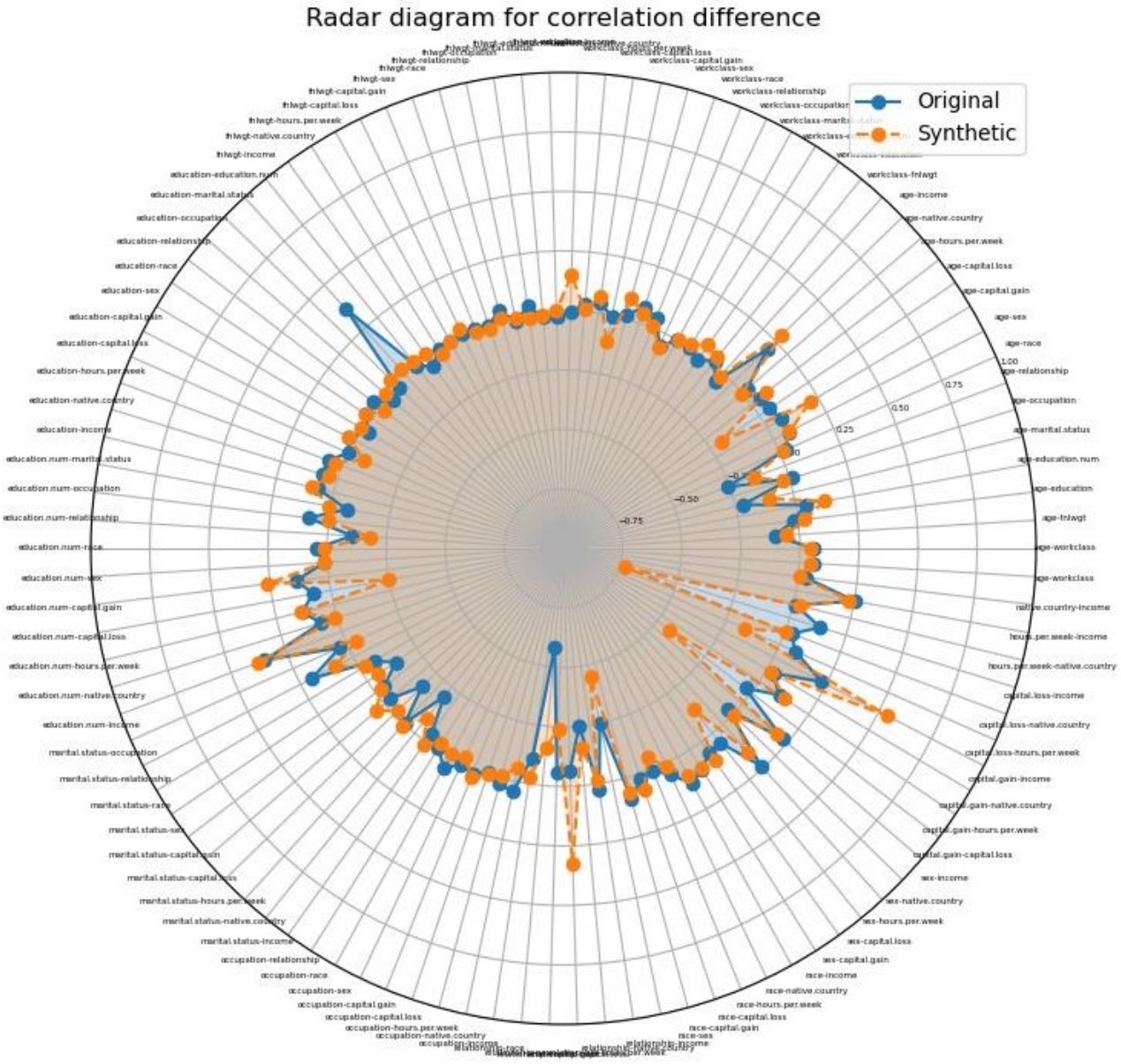
Differences in the performance of synthetic models can be caused by various factors:

- CTGAN is known for generating continuous data and correctly preserving the marginal distributions of variables but may struggle with accurately reproducing complex correlations between variables. Hypothetical example: CTGAN might have difficulties preserving a complex correlation between "monthly salary" and "number of hours worked per week," resulting in a significant deviation from the original correlation.
- SYNTHPOP employs regression-based methods and may perform better for variables with explicit regression relationships. Hypothetical example: SYNTHPOP may better reproduce the regression relationship between "age" and "work experience" in your data, as this relationship can be modeled through linear regression.
- "Fusionstrap" combines features of different methods, attempting to capitalize on the advantages of each. Extreme example: Fusionstrap may achieve balanced performance in some cases but may struggle in situations where data contains complex and non-uniform relationships. Hypothetical example: Fusionstrap may achieve balanced performance in reproducing relationships between "education" and "income" in the case of uniform distributions but may struggle in situations where this relationship is complex and variable.

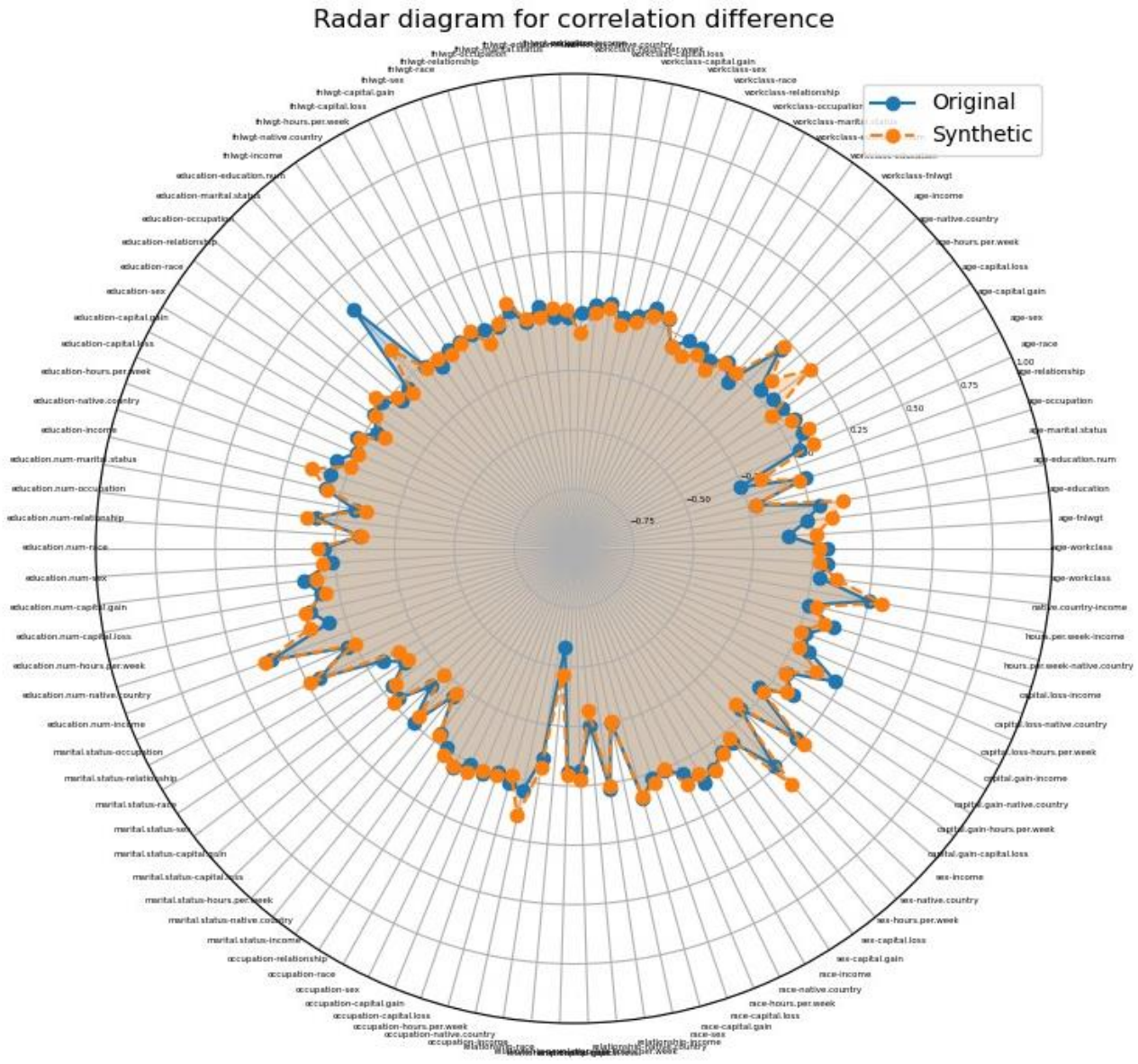
SYNTHPOP might be preferred in situations emphasizing linear or regression relationships, while CTGAN could perform better in cases of complex marginal distributions. Fusionstrap could be a balanced option, but it is essential to evaluate it based on specific data and analysis objectives. In general, interpreting the utility of synthetic data should consider the specific nature of the data and analysis requirements, choosing the method that optimizes the reproduction of significant relationships.

Figure 27

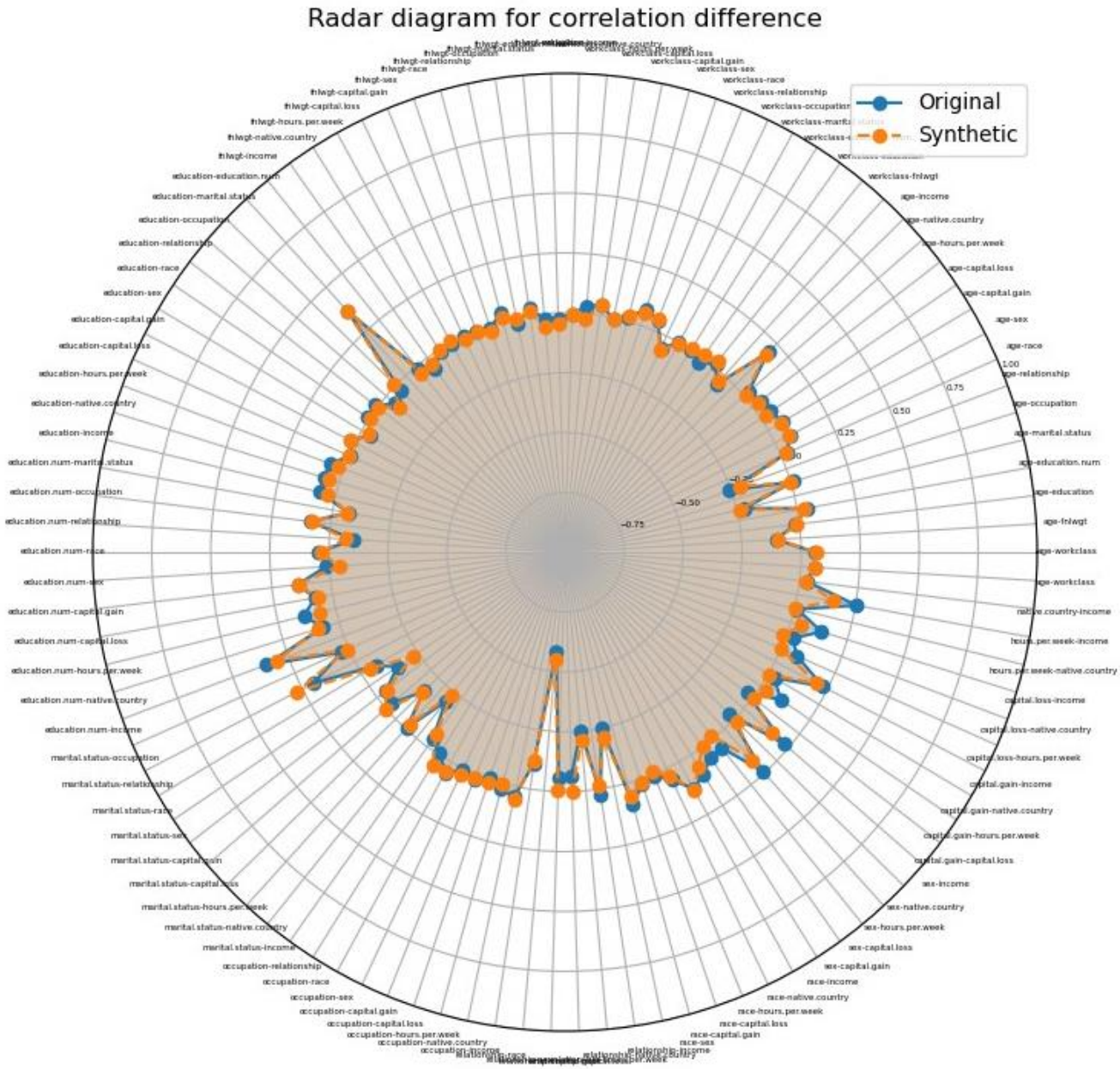
Radar diagram correlation synthetic-original for US census (a. "Fusionstrap"; b. CTGAN; c. SHYNTPOP.)



a.


Radar diagram for correlation difference
Original
Synthetic
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

b.



c.

- ***Diabets Prediction***

Radar charts for the Diabetes Prediction dataset **(Figure 28**) indicate better accuracy in preserving correlations from the original data for the SYNTHPOP method, followed by CTGAN, with "Fusionstrap" showing variations for several correlation pairs. The analysis for this dataset revealed the following observations:

- Value Range **(Appendix B)**:

Minimum: "Fusionstrap": 0.001914897, CTGAN: 0.000963539, SYNTHPOP: 0.000760228
Maximum: "Fusionstrap": 0.441716990, CTGAN: 0.250465857, SYNTHPOP: 0.090552499

- Percentage of Cases with Minimum Values **(Appendix B):**

Fusionstrap: 11%
CTGAN: 19%
SYNTHPOP: 67%

- Interpretation on Variable Groups **(Appendix B)**:

In most variable pairs, SYNTHPOP consistently records the smallest differences, indicating a superior ability to reproduce relationships in the Diabetes dataset. "Fusionstrap," falling within a medium-performance range, has significant advantages in certain scenarios, such as variable pairs "age+hypertension," "gender+smoking_history," "heart_disease+smoking_history," and "hypertension+smoking_history," where it records significantly smaller differences than CTGAN and performs comparably to SYNTHPOP. CTGAN, despite recording larger differences in some cases, may offer reasonable performance, suggesting adaptability to the characteristics of the Diabetes dataset.

- General Conclusion:

SYNTHPOP records the smallest differences in 67% of variables, highlighting better performance in accurately reproducing relationships between variables for the Diabetes dataset. Fusionstrap falls within a medium range, with the smallest differences in 11% of variables. CTGAN, despite recording a high number of variations, demonstrates better performance than Fusionstrap in 19% of variables.

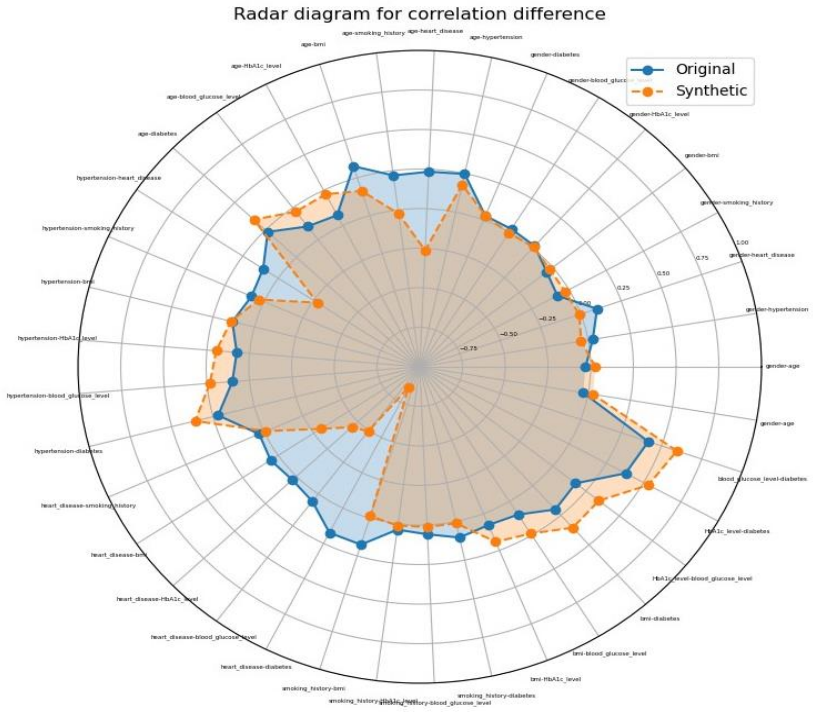
Factors influencing differences in model performance could include:

- Variables related to diabetes may exhibit intricate relationships, and SYNTHPOP, with its regression-based approach, might be more efficient in capturing complex associations.
- Fusionstrap's combination of methods allows adaptability to different types of relationships but may face challenges in scenarios with extremely non-linear or complex correlations.
- CTGAN, known for preserving marginal distributions, might struggle with precisely reproducing specific correlations between variables related to diabetes.
- SYNTHPOP's regression methods may excel in scenarios where explicit regression relationships play a crucial role, such as in diabetes-related predictors.
- Variables related to diabetes, such as "HbA1c_level," "blood_glucose_level," and "diabetes" itself, may have higher importance in the SYNTHPOP model, contributing to its overall superior performance.

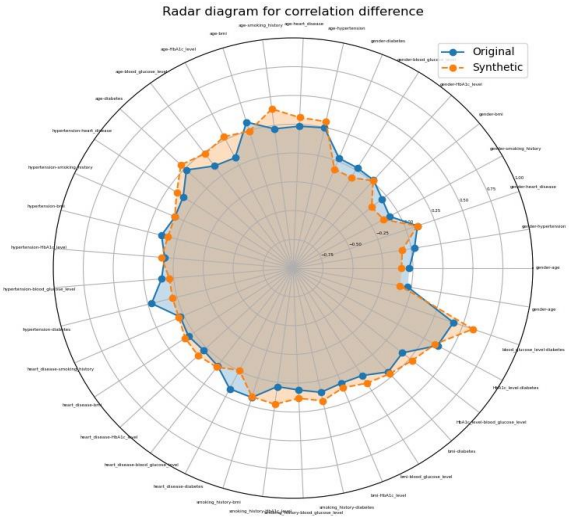
Given the range of differences (0.0007:0.44), we can appreciate that the changes are insignificant for all three methods.

Figure 28

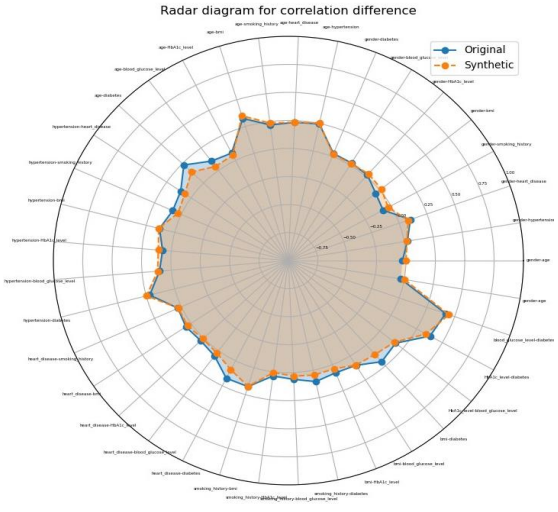
Radar diagram correlation synthetic-original for Diabet Prediction (a. "Fusionstrap"; b. CTGAN; c. SHYNTPOP).



a.



b.



c.

➢ ***AIDS***

The scatterplots for the AIDS data set (Figure 29, Figure 30 and Figure 31) show a uniform distribution of points around the diagonal line of increasing direction, and most of the coefficient values are close to 0 This suggests that the analyzed data pairs did not undergo significant changes in the correlations in the synthetic data compared to the original ones. From the Figure 31, it can be inferred that the SHYNTPOP method has the best accuracy in keeping the correlations from the original set (the points are closely aligned around the diagonal), followed by "Fusionstrap" and CTGAN.

Figure 29

Scatterplot diagram correlation synthetic-original for AIDS with "Fusionstrap"

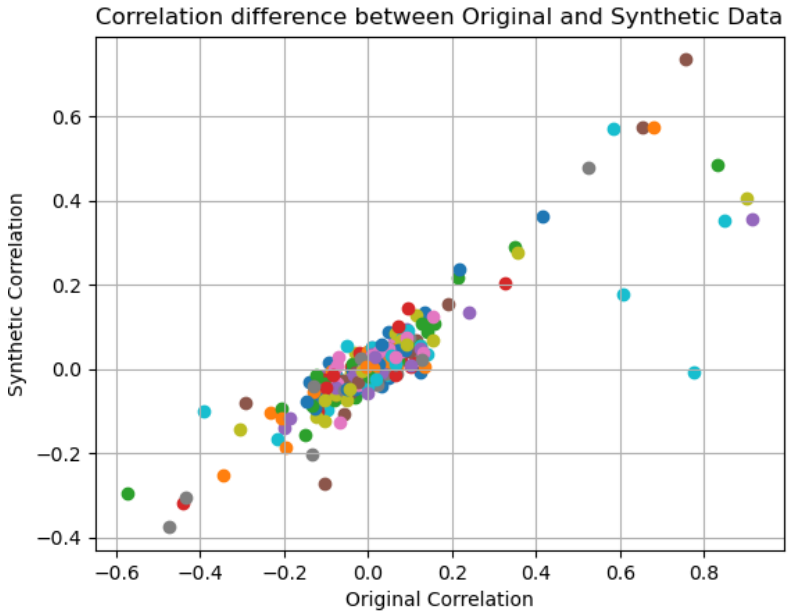


Figure 30

Scatterplot diagram correlation synthetic-original for AIDS with CTGAN

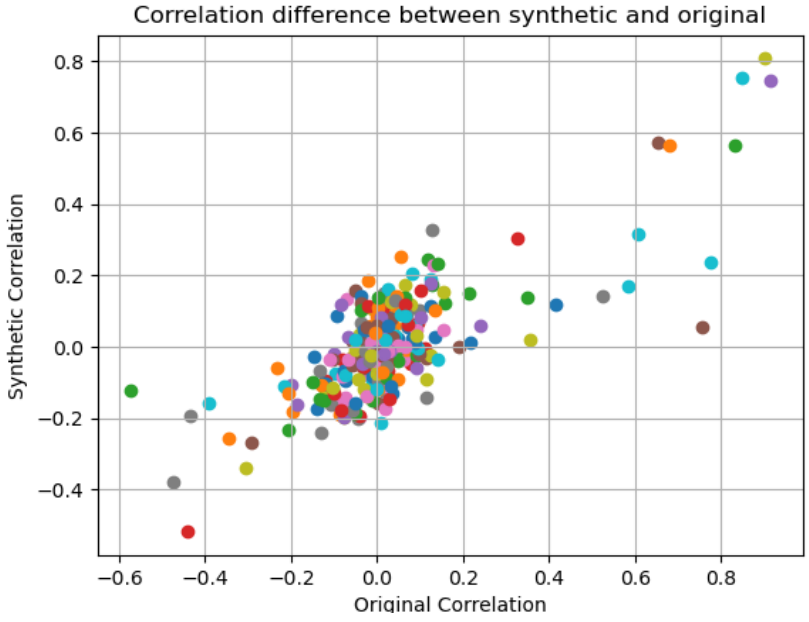
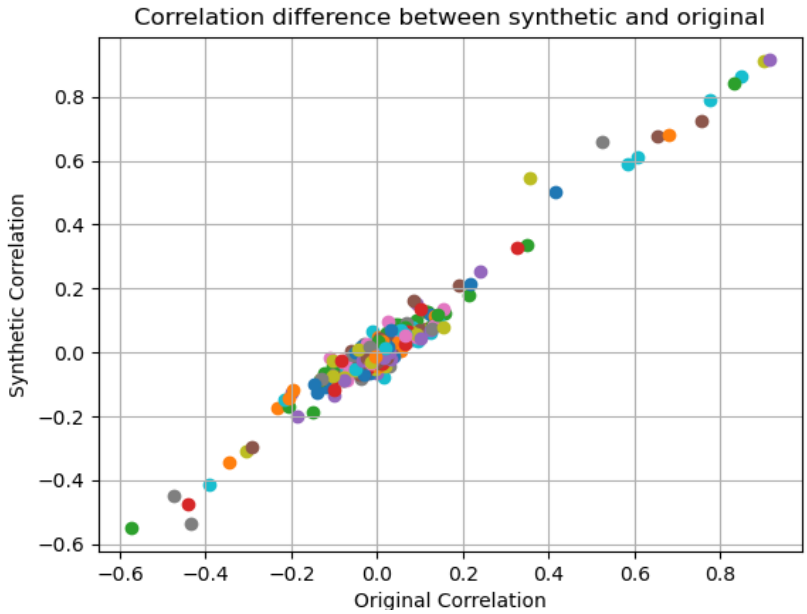


Figure 31
Scatterplot diagram correlation synthetic-original for AIDS with Synthpop



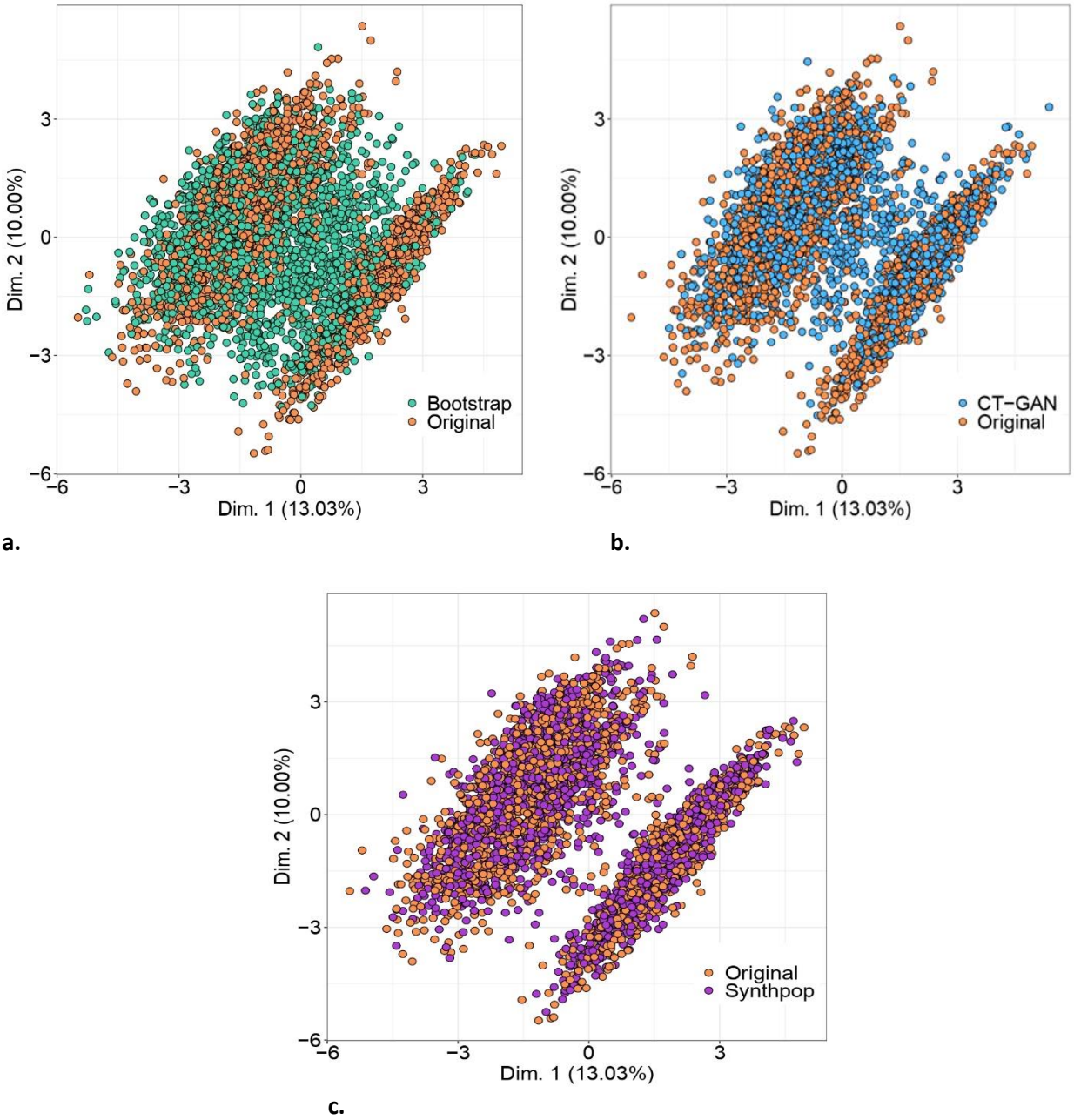
In conclusion, detailed analysis of the correlations in the synthetic data indicates that "Fusionstrap" and SYNTHPOP show good preservation of correlations from the original data, while CTGAN shows some variation. This underlines the promising abilities of the two methods in preserving the complex structures of the original data.

6.3.1.3 Factorial Analysis of Mixed Data (FAMD)

In the evaluation of the AIDS dataset, Factor Analysis of Mixed Data (FAMD) will be used to illustrate and interpret the intrinsic complexity and relationships between the synthetic data generated by various methods. This approach aims at a deeper understanding of the underlying structures of these data. In the analysis process, we will use FAMD to project instances into a Euclidean space, thus avoiding dimensionality challenges by resizing the space and optimizing the computational process.
The algorithm of the method [106] follows the steps:

- Computation of matrices for categorical and continuous variables: For categorical variables, a correspondence matrix is constructed to highlight relationships between categories. In the case of continuous variables, the covariance or correlation matrix is calculated, depending on the nature of the data.
- Matrix decomposition: By applying Correspondence Factor Analysis (AFC) to categorical variables and Factor Analysis (AF) to continuous ones, the latent factors that explain the variation in the data are identified.
- Factor integration: The factors obtained from the AFC and AF analyzes are combined into a single FAMD analysis, with an appropriate weighting for each type of variable.
- Interpretation of factors: Graphical visualization of factors and the contribution of each variable to them facilitates understanding of the relationships between variables and their influence on the underlying structures of the data.

Figure 32 presents comparative results of analyzes based on original data and the three approaches “Fusionstrap” (a), CTGAN (b) and SYNTHPOP (c).



Note. a. FAMD projections of AIDS data generated using "Fusionstrap" within the original data space (original data represented by orange points, "Fusionstrap" data depicted by green points). b. FAMD projections of AIDS data generated with CTGAN within the original data space (original data denoted by orange points, CTGAN data represented by blue points). c. FAMD projections of AIDS data generated using SYNTHPOP within the original data space (original data marked by orange dots, SYNTHPOP data shown by purple dots).

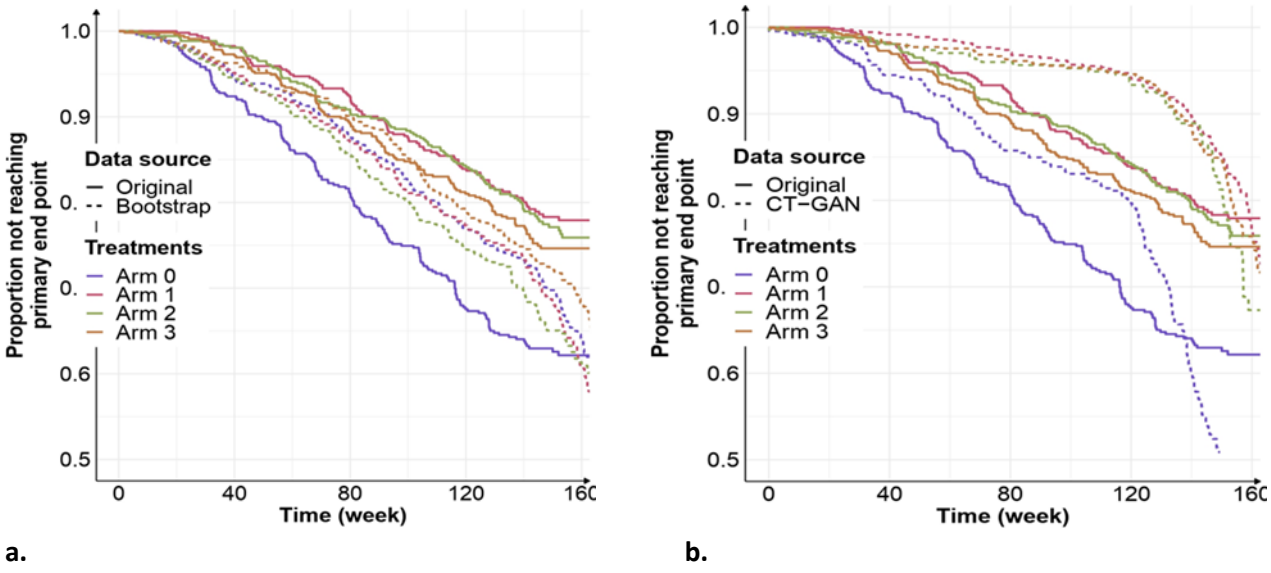
All three methodologies maintain the statistical significance of the original AIDS dataset, as evidenced by the superimposition of data points from the three methods onto the same space constructed from the original observations. This alignment indicates a utility level comparable to the original dataset. Upon scrutinizing the "Fusionstrap" method's performance concerning Factor Analysis of Mixed Data (FAMD) on

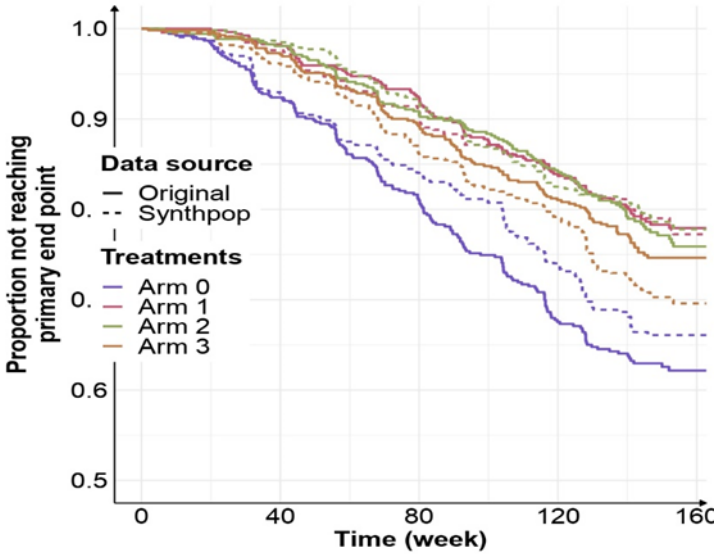
the AIDS dataset and juxtaposing it with Synthpop and CT-GAN methods, it becomes apparent that "Fusionstrap" excels in preserving the dataset's essential features. The FAMD projections for the data generated by "Fusionstrap" exhibits a complete overlap with the original set, encompassing outliers, emphasizing its superior utility in this particular experimental context. This noteworthy agreement underscores the capability of the "Fusionstrap" method to accurately replicate the essential features of the dataset in FAMD analysis, emphasizing its efficacy in preserving and faithfully reproducing the data distribution when compared to the Synthpop and CTGAN methods.

6.3.1.4 Checking fidelity with "Avatar" statistics: Survival Curve and Hazard Ratio

To assess the capacity to preserve the utility of the original AIDS dataset, we examined how the values of two predictive statistics, namely the "survival curve" and "hazard ratio," were sustained within the synthetic datasets.
The survival curve (Survival Curve) represents the probability of survival to a certain point in time for a group of subjects or patients. This curve is made by calculating the proportion of subjects that survived to each specific time point (survival rate). Based on the survival rates, the survival curve is constructed.
The Hazard Ratio serves as an indicator of the disparity in the occurrence of events between two groups, as observed in our case through a comparison between synthetic and original data. It gauges the rate at which events unfold in one group relative to the rate in the other group. To calculate the hazard ratio, we will determine the hazard rate for the original data set and for each synthetic data set obtained through the three approaches (the ratio between the number of events and the time at risk). Time at risk is the length of time that study subjects are at risk of experiencing the event (death). Then we will calculate the ratios between the hazard rates for the synthetic data and the hazard rate for the original data. If the hazard ratio is equal to 1, it means that there is no significant difference between the compared data sets in terms of the spread of events. If the hazard ratio is greater or less than 1, it indicates a significant difference in the spread of events between the groups.
Figure 33 compares the four-treatment survival curves calculated on the synthetic data generated with "Fusionstrap" (a), CTGAN (b), SYNTHPOP (c) and the original SIDA data set. The survival curves for the four treatment arms exhibit a more pronounced overlap between the data generated by "Fusionstrap" and SYNTHPOP (dotted line) and the original data (solid line).

Figure 33
Comparative results: a. Survival curves – "Fusionstrap"; b. Survival curves – CTGAN; c. Survival curves – SYNTHPOP.





c.

Figure 34 and Figure 35 demonstrate that the main results of the study regarding the survival curve in time and the hazard ratio remained unchanged for all three approaches, with small changes in the case of the survival curve for CTGAN.

Figure 34

Comparative survival curves in time

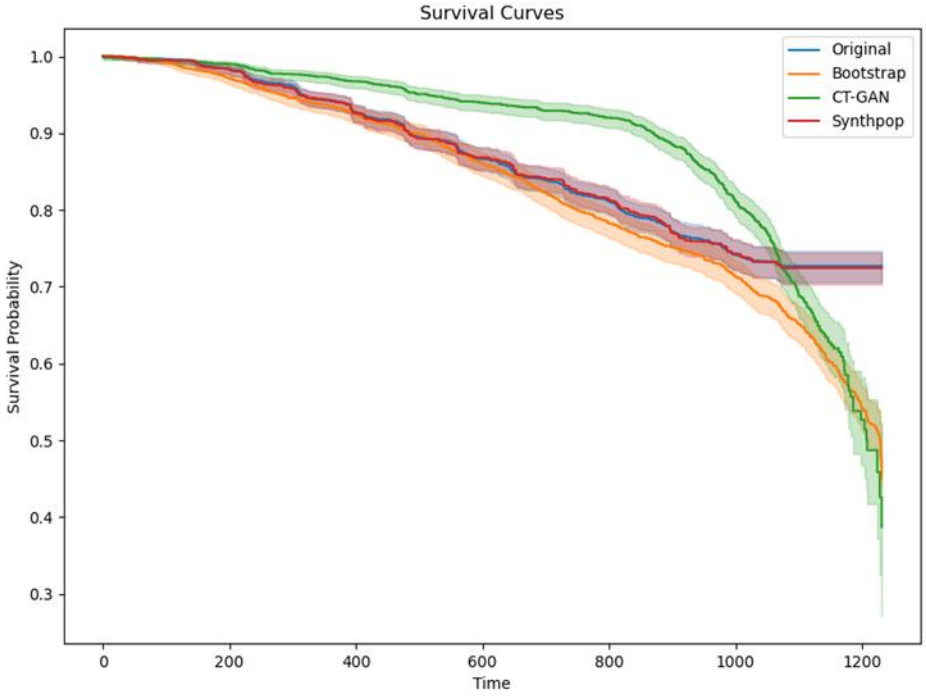


Figure 36 indicates the comparison between the three approaches for the P-value for the hazard ratio (Hazard Ratio). This statistical metric assesses the statistical significance of the difference between the hazard rates of two groups, as observed in our case between the synthetic and original datasets. In other words, p-values help us decide whether any observed differences are the result of random variation or represent significant differences. In the framework of our analysis and with a selected significance level of 0.05, a p-value exceeding this threshold (e.g., 0.8 for "Fusionstrap") implies that the observed differences could potentially arise from random variation. Consequently, there is insufficient evidence to assert the

significance of these potential differences. Table 10 presents the computed values for hazard ratio and p-value.

Table 10
Values for hazard ratio

Method	Hazard Ratio	p-value
Original	1.009496	0.079344
Fusionstrap	0.998804	0.786500
CTGAN	0.999295	0.893157
Synthpop	1.001635	0.753656

Figure 35
Comparative Hazard Ratio

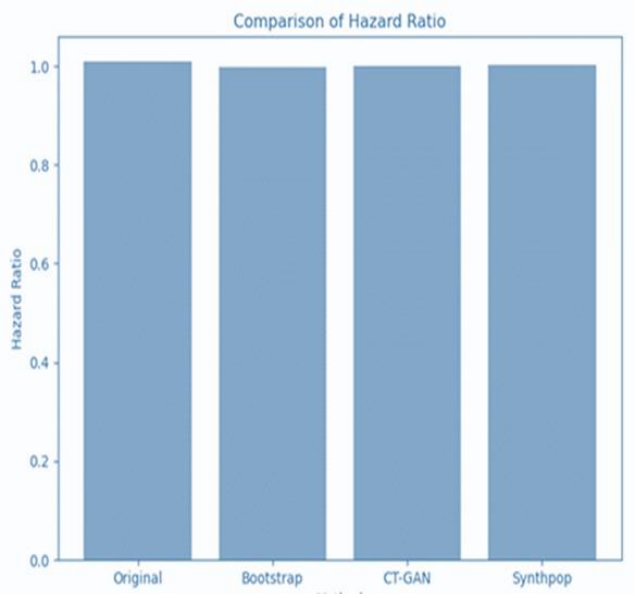
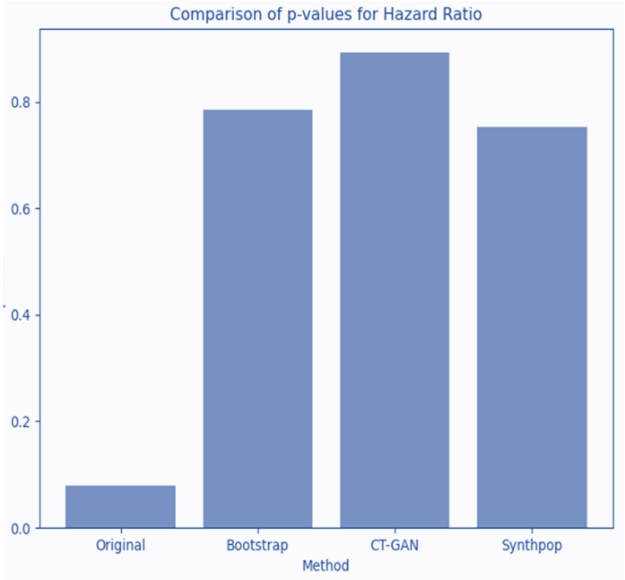


Figure 36
Comparative p-values for Hazard Ratio



When considering the assessment of predictive hazard ratio statistics and survival curve preservation in synthetic data, we broadened our perspective by examining analogous research on the AIDS dataset. This prior study utilized a synthesis method known as Avatar [94]. The main results of this study [94] show a similar effectiveness of "Fusionstrap" to that of avatar. For both approaches, the effectiveness of arm 1 in the survival curve, when comparing CD4 T-cell counts over time, surpassed that of arm 0. Specifically, for "Fusionstrap," the hazard ratio (HR) was 0.998804 compared to the original HR of 1.009496, and the p-value was 0.786500 versus the original p-value of 0.079344. In contrast, for Avatar, the hazard ratio was 0.40 compared to the original HR of 0.49, and the p-value was 1.47e-11 versus the original p-value of 1.22e−08. These findings support the consistency and relevance of the “Fusionstrap” method in generating

accurate and useful synthetic data, providing a solid foundation for its applicability in similar research areas.

Summary:
Within this section, we delved into an exploration and assessment of the effectiveness of synthetic data produced by the "Fusionstrap" framework in comparison to alternative methods such as CTGAN and SYNTHPOP. Our goal was to deeply understand the performance of these methods in the context of generating synthetic data, as well as how they were able to preserve the essential features and relationships from the original datasets. In evaluating the correlations, we found that “Fusionstrap” achieved results comparable to those of the CTGAN and SYNTHPOP methods. The points in the scatterplot and the polygons in the radar plot indicated that “Fusionstrap” was able to effectively maintain relationships between variables without introducing significant bias. In addition, we used Factor Analysis of Mixed Data (FAMD) to examine in more detail the complex structures of the synthetic data and how these were modeled by the three methods. Our results highlighted “Fusionstrap”'s ability to preserve the essence of the original data, contributing to a deeper understanding of the relationships between variables. In conclusion, the “Fusionstrap” framework has demonstrated robust performance in generating synthetic data, maintaining essential features and relationships from the original datasets. This suggests that “Fusionstrap” is a promising option for generating synthetic data in the context of research and data analysis. In our evaluation, the “Fusionstrap” framework achieved the best results compared to the other methods (CTGAN and SYNTHPOP) in terms of preserving essential features and relationships between variables in the AIDS (HIV infection) dataset. This implies that "Fusionstrap" effectively captured the distributions and relationships within this dataset without introducing notable bias, indicating superior performance compared to other methods, especially on smaller and more sensitive datasets. It's crucial to acknowledge that the efficacy of the "Fusionstrap" method can fluctuate based on the unique characteristics of each dataset and the specific demands of the analysis.

6.3.2 Evaluation of privacy

Safeguarding the privacy of personal and sensitive information is a pivotal challenge when utilizing and exchanging data. In an ever-evolving digital society where data is used for decision-making and research, protecting private information is crucial to maintaining trust and complying with ethical and legal norms. To address this issue, we conducted a comparative privacy assessment of synthetic data generated by “Fusionstrap”, CTGAN, and SYNTHPOP, on three distinct datasets: US Census, Diabetes Prediction, and HIV Infection (AIDS). Through the privacy benchmarking, we aimed to determine whether synthetic data generated by “Fusionstrap” can provide a viable and safer alternative for data use and sharing compared to traditional methods. For this evaluation, the method based on data holdout and the evaluators described in detail in Section 3.5 of this thesis were used. The holdout evaluation was done by dividing the synthetic data sets into holdout sets (20% of the number of records) and training sets (80% of the number of records).
The results of the comparative evaluation on the three data sets are presented in Tables 11, Tables 12 and Tables 13**,** highlighting the aspects in which “Fusionstrap” proved to be more promising in terms of keeping integrity and confidentiality of information.

6.3.2.1 Holdout method and evaluators

Table 11
US census data

EVALUATORS	“FUSIONSTRAP”	CTGAN	SYNTHPOP
Statistical evaluators			
KS test	0.53	0.74	0.99
CS test	0.53	0.97	1.00
Identical match			
Ratio in holdout	0	0	0

Ratio in shyntetic	0	0	0
DCR test			
Shyntetic&Real	8.69	5.83	3.13
Holdout&Real	8.69	5.82	3.14
NNDR test			
Shyntetic&Real	0.83	0.53	0.22
Holdout&Real	0.20	0.20	0.20
Data detection			
Logistic detection	0.005	0.04	1.00
Random forest	0.004	0.04	0.94

Table 11 provides a comparative assessment of the privacy risk associated with using synthetic data generated by "Fusionstrap", CTGAN, and SYNTHPOP from the US Census set. The evaluations were conducted using several evaluation metrics, each providing a different perspective on the effectiveness of each method in protecting data privacy. The KS and CS tests evaluate the similarity between synthetic and real distributions. In the context of protecting privacy, the goal is to obtain synthetic data that has a distribution as close as possible to the real one, without revealing sensitive or personal information. a smaller value indicates less divergence between the distributions, suggesting that the synthetic data are less discernible from the real data. Therefore, the lower values for "FUSIONSTRAP" in these tests (KS= 0.53. CS= 0.53) suggest that this method provides synthetic data with a better similarity to real data while providing greater privacy protection than CTGAN and Synthpop.

Zero values for Ratio in holdout and Ratio in synthetic suggest that there are no identical records between the real and synthetic datasets for either method. Lack of identical matching is often a positive aspect in generating synthetic data, as it shows that individual records from the original data set have not been recreated in detail. Therefore, the obtained results indicate a good practice in terms of privacy protection, since there is no identical data between the real and synthetic datasets.

For DCR test and NNDR test, "Fusionstrap" shows higher values for the two comparisons (Synthetic&Real and Holdout&Real), signifying a better fit of the synthetic data to the real data compared to the other methods. Regarding the data detection tests (Logistic detection and Random forest), the results suggest that the synthetic data generated by "Fusionstrap" has the lowest risk of being detected as synthetic compared to that generated by CTGAN and SYNTHPOP.

In conclusion, for this dataset, based on the analyzed results, "Fusionstrap" appears to provide better privacy protection compared to the other two methods, with better values in most of the privacy risk assessment tests.

Table 12

Diabetes Prediction data

EVALUATORS	"FUSIONSTRAP"	CTGAN	SYNTHPOP
Statistical evaluators			
KS test	0.69	0.86	0.99
CS test	0.23	0.82	1.00
Identical match			
Ratio in holdout	0.02	0.02	0.02
Ratio in shyntetic	0.00	0.00	0.01
DCR test			
Shyntetic&Real	3.05	2.21	1.61
Holdout&Real	3.16	2.19	1.61
NNDR test			
Shyntetic&Real	0.43	0.80	0.36

Holdout&Real	0.33	0.12	0.33
Data detection			
Logistic detection	0.20	0.45	1.00
Random forest	0.20	0.46	0.99

The comparative privacy risk analysis of the "Diabetes Prediction" dataset (Table 12) reveals significant differences between the three synthetic data generation methods: “Fusionstrap”, CTGAN and SYNTHPOP. Following statistical tests such as KS test and CS test, we notice that the values associated with “Fusionstrap” are lower compared to those of CTGAN and SYNTHPOP methods. This may suggest that the synthetic data generated by “Fusionstrap” more closely resembles the real data in terms of distributions. Looking at the goodness-of-fit tests, we see that all three methods show some level of differentiation between holdout and synthetic data. However, “Fusionstrap” seems to keep these differences to a smaller level, indicating a potentially more effective approach to protecting privacy. The results for DCR test and NNDR test show that “Fusionstrap” exhibits a better fit between synthetic and real data compared to CTGAN and SYNTHPOP. This indicates that the data produced by "Fusionstrap" possesses a superior capacity to mirror the authentic characteristics of the dataset. When considering data detection tests, "Fusionstrap" seems to hold notable advantages. Lower values for Logistic detection and Random Forest indicate a lower probability that synthetic data generated by “Fusionstrap” will be detected as artificial compared to the other two methods.
In conclusion, the analysis on the "Diabetes Prediction" dataset reveals that “Fusionstrap” presents a more robust approach to privacy risk management compared to CTGAN and SYNTHPOP.

Table 13
AIDS data

EVALUATORS	“FUSIONSTRAP”	CTGAN	SYNTHPOP
Statistical evaluators			
KS test	0.80	0.82	0.94
CS test	0.58	0.92	0.99
Identical match			
Ratio in holdout	0	0	0
Ratio in shyntetic	0	0	0
DCR test			
Shyntetic&Real	12.23	9.50	8.34
Holdout&Real	12.26	9.42	8.37
NNDR test			
Shyntetic&Real	0.67	0.54	0.47
Holdout&Real	0.46	0.46	0.46
Data detection			
Logistic detection	0.06	0.35	0.67
Random forest	0.06	0.24	0.70

Examining the risk of privacy compromise on the "AIDS" data set (Table 13), we notice that, also in the case of this data set, KS test and CS present lower values for “Fusionstrap” compared to CTGAN and SYNTHPOP. This suggests that the synthetic data created by “Fusionstrap” has a better match with the real data in terms of their distributions. The results from the DCR and NNDR tests show that “Fusionstrap” is closer to the authentic data compared to the other methods. This indicates that the data generated by “Fusionstrap” better captures the essential features of the original data set. In terms of data detection analysis, the lower values for Logistic detection and Random Forest suggest that the synthetic data obtained by “Fusionstrap” is less likely to be detected as artificial, in contrast to the data generated by CTGAN and SYNTHPOP.
In conclusion, the privacy risk analysis on the "AIDS" dataset confirms that “Fusionstrap” stands out as more robust in protecting privacy compared to the other two methods.

6.3.2.2 Re-identification attack

This experiment will perform a re-identification risk assessment on the synthetic datasets generated by the three methods: Bootstrap, CTGAN and SYNTHPOP. The purpose of the experiment is to determine how often a distance-based linking attack can lead to the correct re-identification of the individual in the synthetic data set.
The comparative evaluation between the three methods will go through the following steps:

- The original datasets and the three synthetic datasets (Bootstrap, CTGAN and SYNTHPOP) are defined as inputs.
- The identification attributes, crucial for re-identification purposes, are explicitly determined. These variables encompass the individual's attributes in the dataset that are prone to be associated with their identity.
- The target variable for re-identification is defined. This represents the attribute that is desired to be correctly re-identified in the synthetic dataset.
- A dictionary is initialized to store the correct re-identification rate results for each method.
- Iterate through each individual record in the original data set. For each individual, the distances between its attributes and the attributes of each individual in each synthetic dataset (Bootstrap, CTGAN and SYNTHPOP) are calculated. The synthetic individual with the smallest distance from the original individual is found.
- The target variable of the synthetic individual is compared with that of the original individual. If these target variables are identical, then a correct re-identification has been achieved.
- Determine the accurate re-identification rate for each method by dividing the number of correct re-identifications by the total number of attempts at re-identification.

The results in Table 14 indicate the re-identification accuracy for 100 attempts in the case of each method (Bootstrap, CTGAN and SYNTHPOP). This accuracy refers to how often a correct re-identification of the individual from the original dataset occurred in the generated synthetic dataset. The target variables for re-identification were: "income" for US census, "diabetes" for Diabetes Prediction and "offtrt" (the variable that provides information on whether or not a patient came off antiretroviral treatment (ART) within a certain interval of time) for the AIDS set.

Table 14
Re-identification accuracy

Re-identification accuracy	"Fusionstrap"	CTGAN	SYNTHPOP
US census	0.31	0.68	0.67
Diabet Prediction	0.70	0.90	0.93
AIDS	0.61	0.64	0.36

The data presented in Table 14 regarding re-identification accuracy for each method ("Fusionstrap", CTGAN and SYNTHPOP) on the three datasets indicate the level of risk associated with exposing sensitive data. Re-identification accuracy refers to how often the individual from the original data set was correctly identified in the generated synthetic set. The higher the re-identification accuracy, the higher the risk of exposing sensitive data. Comparing the results for each dataset, we can see that for the US census dataset, "Fusionstrap" shows the lowest re-identification accuracy (0.31), followed by SYNTHPOP (0.67) and CTGAN (0.68). This indicates that, in this context, "Fusionstrap" can provide better protection against the risks of re-identification and exposure of sensitive data. Regarding the Diabetes Prediction dataset, CTGAN and SYNTHPOP record higher accuracies (0.90 and 0.93, respectively), indicating an increased risk of re-identification and exposure of sensitive data for these two methods. "Fusionstrap" shows lower accuracy (0.70), suggesting less exposure of sensitive data compared to the other methods. For the AIDS dataset, the re-identification accuracy is lowest for SYNTHPOP (0.36), followed by "Fusionstrap" (0.61) and CTGAN (0.64). Thus, SYNTHPOP can provide better protection against the risk of re-identification and exposure of sensitive data in this context.

In conclusion, analyzing these results, we can see that “Fusionstrap”, in most cases, presents a lower re-identification accuracy compared to the other methods, indicating a lower exposure of sensitive data.

6.3.3 Analysis of class imbalances

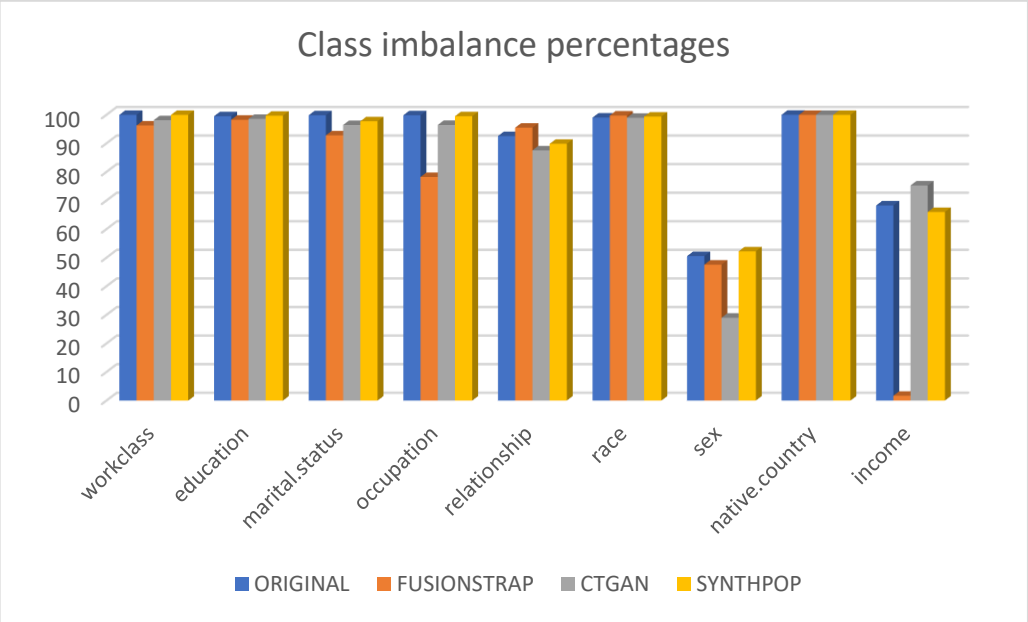
In this segment of the analysis, we focused on evaluating the percentages of class imbalances in the synthetic data sets obtained by means of the “Fusionstrap” method. The key objective of this step was to evaluate the proficiency of “Fusionstrap” in correcting class imbalances compared to the original datasets. We also evaluated how this methodology compares to the CTGAN and SYNTHPOP alternatives in addressing this aspect of synthetic data generation.

The results obtained for each variable for the three data sets analyzed are presented in Figure 37, Figure 38 and Figure 39.

For the US census data set (Figure 37), we observe that “Fusionstrap” was able to significantly reduce class imbalances for most variables, while still maintaining an acceptable proportion of the class distribution. For example, variables such as "workclass", "education", "marital.status" and "occupation" saw notable improvements in handling inequities between the corresponding classes. This suggests that “Fusionstrap” had a positive impact on obtaining balanced synthetic data, better reflecting the original distribution. In comparison, the CTGAN and SYNTHPOP methods produced mixed results, sometimes maintaining the original class proportions but also generating significant inequities in certain variables.

Figure 37

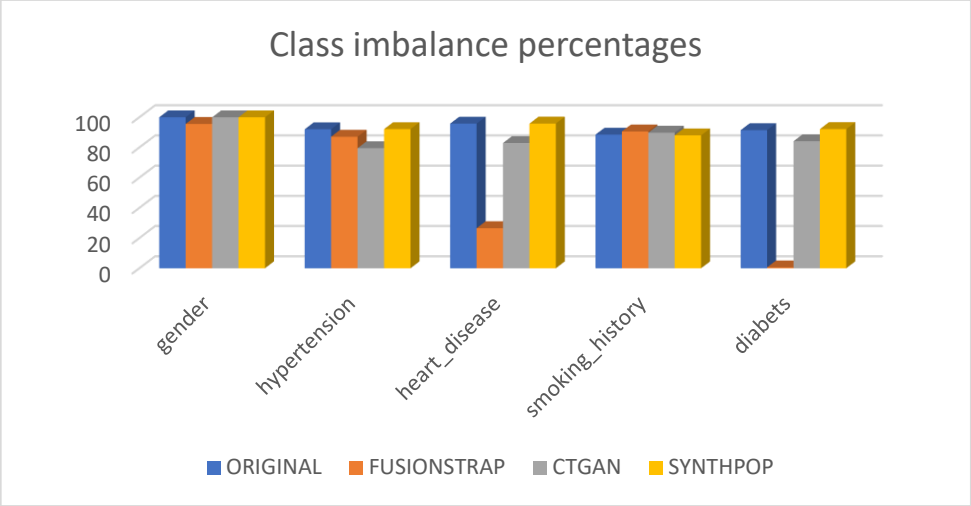
Class imbalance percentages for US census



In the case of the Diabetes Prediction set (Figure 38), “Fusionstrap” had a significant effect in reducing class imbalances for the variables: "gender", "hypertension" and "smoking_history". These results suggest that the “Fusionstrap” method helped balance the distribution of classes for these variables so that they better reflect the original structure of the data. While the CTGAN method achieved improvements in class balancing for some variables, SYNTHPOP showed mixed performance. It is crucial to highlight that, concerning the "diabetes" variable, "Fusionstrap" successfully mitigated the class imbalance, significantly reducing the percentage from 91.45% to only 0.6%. This exemplifies the efficacy of the method in addressing class inequities within the Diabetes dataset Prediction.

Figure 38

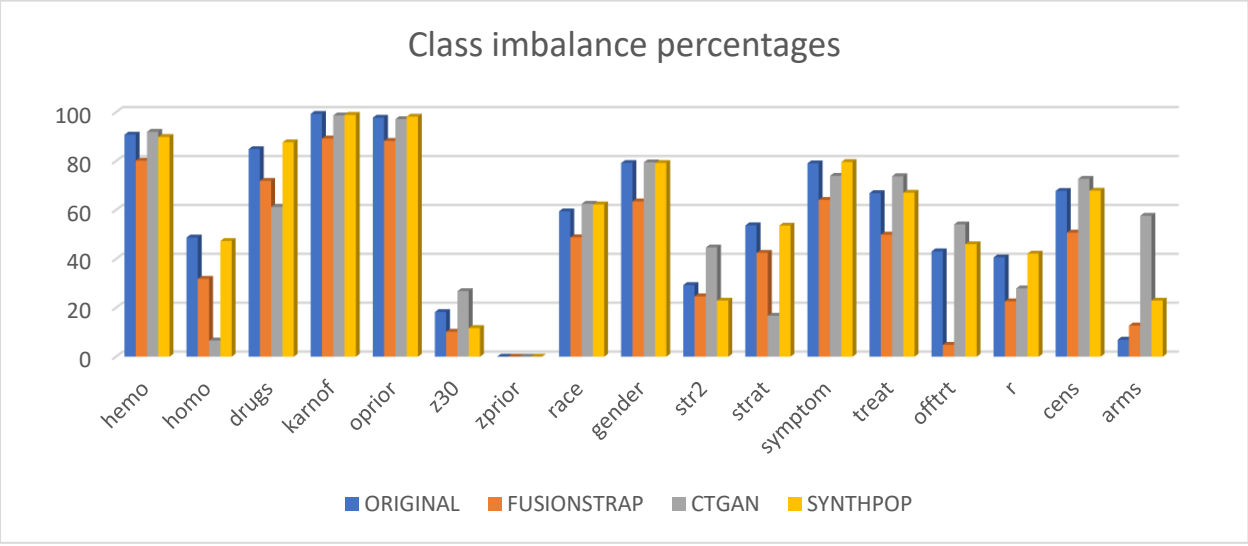
Class imbalance percentages Diabetes Prediction



In Figure 39, we notice that in the case of variables "homo", "z30", "zprior", "str2" and "offtrt", the “Fusionstrap” method managed to bring a significant balance between the classes compared to the original set, demonstrating its effectiveness in dealing with these class imbalances. Also, for the variables "hemo", "gender", "strat", "treat" and "r", “Fusionstrap” obtained significant improvements in balancing the class distribution. However, the method recorded a slight decrease in class balance for the "race" variable. Compared to the CTGAN and SYNTHPOP methods, “Fusionstrap” showed better performance in handling class inequities for most variables in the AIDS dataset.

Figure 39

Class imbalance percentages for AIDS



By analyzing the percentages of class imbalances within each data set (US Census, Diabetes Prediction, and AIDS), we can conclude that “Fusionstrap” performed superiorly in treating and reducing class inequities compared to the other methods. This suggests that “Fusionstrap” manages to preserve a better distribution of classes, thus ensuring a more balanced representation of the data. “Fusionstrap” thus achieves one of the purposes for which it was created, namely, to resolve class imbalances in data sets by generating synthetic data.

Chapter 7
Conclusion

This chapter presents a summary of the “Fusionstrap” framework, highlighting its key aspects, strengths, limitations, and potential directions for further development.

7.1 Summary

This thesis addressed a number of key issues related to the generation and evaluation of synthetic data to improve its quality and utility in the context of class imbalances and data privacy protection. The primary objective of this study was to create and assess a framework named "Fusionstrap," drawing comparisons with conventional methods like CTGAN and SYNTHPOP. In the first part of the thesis, we explored the problem of class imbalances in datasets and analyzed their impact on machine learning algorithms. We have shown that class imbalances can lead to unfair decisions and that generating synthetic data can be a solution to correct this problem. Next, we presented the “Fusionstrap” framework, an integrated data processing and synthesis system developed in response to the challenges posed by class imbalance in datasets. This framework addresses several essential aspects of data processing, including rigorous preprocessing, advanced synthetic data generation using Bootstrap to handle class imbalances, and postprocessing techniques such as removing outliers and applying Bootstrap rotation to obtain a unique set of synthetic data. Comprehensive assessment of the utility and privacy of synthetic data, compared to other data synthesis methods, completes this framework. The diagram of relevant components (Figure 16) provides a visual illustration of the architecture and functionality of the “Fusionstrap” system, showing how each stage contributes to improving data quality and ensuring privacy, in a context of effective class imbalance management. To assess the efficacy of "Fusionstrap," a set of experiments was conducted using three datasets, and the outcomes were juxtaposed with those derived from CTGAN and SYNTHPOP. The examination covered not only the statistical faithfulness of the synthetic data but also its capability to rectify class imbalances. The use of "Fusionstrap" brings significant advantages in the field of synthetic data generation, notable for the following aspects:

- ***Comprehensive Approach:*** "Fusionstrap" is not limited to synthetic data generation, but provides a complete solution for data management and synthesis. It starts with rigorous preprocessing, continues with the generation of synthetic data, and includes postprocessing techniques to ensure data quality and confidentiality. It also includes a full assessment of the utility and privacy of the synthetic data obtained. This assessment is essential to ensure that the synthetic data is relevant to the analysis tasks and complies with confidentiality requirements.
- ***Handling Class Imbalances:*** One of “Fusionstrap”'s distinguishing features is its efficient handling of class imbalances. By using the Stratified Bootstrap technique, this framework generates synthetic data that balances the class distribution, thus making the datasets more suitable for training and evaluating machine learning models.
- ***Flexibility and Customization:*** “Fusionstrap” allows customization of the parameters and methods used to generate synthetic data. This level of flexibility makes this framework suitable for various data types and application scenarios.
- ***Reproducibility:*** “Fusionstrap” is designed to be reproducible. This means that by applying the same methodology and parameters within Fusionstrap, the results obtained will be constant and consistent. In other words, if the same initial data set and the same settings of "Fusionstrap" are used repeatedly, the synthetic data generated is expected to be similar or identical. This aspect provides confidence that the process of generating synthetic data is reliable and can be consistently replicated to obtain the same results in various iterations of the experiment or analysis.
- ***Privacy Management:*** “Fusionstrap” integrates post-processing techniques, such as removing outliers and applying Bootstrap rotation, to protect the privacy of the original data.

- ***Benchmarking:*** "Fusionstrap" provides a solid basis for comparing the quality of synthetic data against other data synthesis methods, helping to identify the most suitable method for a given scenario.

Overall, "Fusionstrap" is a comprehensive and efficient solution for handling class imbalances in data used in data analysis and machine learning. This framework adds value in improving data quality and ensuring confidentiality, thus facilitating decision-making and research processes in various fields. After analyzing the results of the experiments, the conclusions highlight that Fusionstrap can be successfully applied in areas such as health data analysis, epidemic forecasting, as well as in social and economic analysis, proving to be effective especially in the case of small data sets up to mediums (2.000-50.000 records).

7.2 Answers to the research questions

In this research, we addressed the central question of the study: ***"Can Fusionstrap improve the quality of synthetic data over known methods such as CTGAN or SYNTHPOP?".*** Considering the challenges encountered in generating synthetic data, the research objectives consisted of evaluating the quality of the synthetic data, examining the effectiveness in handling class imbalances, and comparing the overall performance of Fusionstrap with CTGAN and SYNTHPOP. Next, we will detail the answers to these questions and highlight key takeaways from Fusionstrap's analysis.

RQ1: To what extent can „Fusionstrap" ensure the utility of the data generated?

RQ1.1: How can the utility of synthetic data be measured?

Assessing the usefulness of synthetic data was an essential aspect in ensuring the fidelity and relevance of this artificially generated data. The importance of this process lies in the ability to validate the performance of synthetic data generation methods, ensuring that they accurately reproduce key features of the original data sets and are thus usable in various contexts, from scientific research to business decision making.
The following selection criteria were considered:

- To evaluate the similarity of the distributions between the synthetic and real data, we opted for the Hellinger evaluator based on previous studies that highlighted the effectiveness of the Hellinger distance in evaluating the similarity between probability distributions [61]. This recommends it as a relevant metric for measuring the usefulness of synthetic data.
- Assessing correlations is crucial to ensure that relationships between variables are preserved in synthetic data. We chose to measure the correlation between variables using the Pearson Correlation Coefficient and Cramer's V Correlation given the specific advantages of these metrics in the context of generating synthetic data. The Pearson Correlation Coefficient is applicable to continuous variables and furnishes details about the direction and strength of the linear relationship between them [62]. At the same time, Cramer's V Correlation is effective for categorical variables, providing a measure of the strength of association between them [63]. Radar diagrams and scatter plot diagrams were chosen to represent the results of correlation evaluation as these visualization methods provide an intuitive and comprehensive perspective on how the relationships between variables are preserved in the synthetic data compared to the real ones [63]. Radar diagrams allow differences and similarities to be observed in a visual way, highlighting areas where the synthetic data approaches or deviates from the actual distribution. On the other hand, scatter plot diagrams provide a detailed way of viewing the distribution of points relative to the regression line, making it easier to identify variations between synthetic and real data.
- For a deeper understanding of the relationships between the variables, the Factor Analysis of Mixed Data (FAMD) method was additionally used (Section 6.3.1.3) recognized from previous studies as a powerful method to examine and interpret complex data structures [106].
- To investigate in depth how the essential features of the original datasets are preserved, it is necessary to evaluate their representative statistics. In the present research, we chose to evaluate two essential statistics of the AIDS data set, namely the survival curve and the hazard ratio (Section 6.3.1.4). This choice was based on the existence of a similar study carried out by the Avatar method

[94], which allows comparing the results with the results of other research carried out on the same data sets.

RQ1.2: What is the utility level of the "Fusionstrap" framework compared to other data synthesis methods (CTGAN and SYNTHPOP)?

To evaluate the utility data generated with "Fusionstrap" framework compared to other data synthesis methods, we applied various metrics and evaluation methods. The following results show that the performance of "Fusionstrap" is competitive or even superior to the performance of the CTGAN and Synthpop methods in terms of similarity of the synthetic data to the original data (Section 6.3.1):

- "Fusionstrap" obtained smaller Hellinger distances for certain variables, indicating a greater similarity between the synthetic and real distributions.
- "Fusionstrap" preserved the correlations between variables better than the two methods, a fact highlighted in radar plots and scatterplots.
- Using Factorial Analysis of Mixed Data (FAMD), we found that "Fusionstrap" modeled the complex structures of synthetic data more efficiently.
- Hazard ratio and survival curve analysis showed that Fusionstrap preserved crucial statistics, which is essential in the context of critical datasets such as that associated with diabetes prediction.

The better performance of "Fusionstrap" in creating data with distributions closer to the distributions of the real data compared to the analyzed methods can be justified by the specific characteristics of the method and the way it approaches data synthesis. Here are some possible explanations:

- "Fusionstrap" is based on the concept of Gaussian Copula, which is effective in modeling complex distributions. This approach may be more appropriate in generating synthetic distributions that are closer to the real ones.
- It is possible that the default settings or configurations chosen for "Fusionstrap" better match the variables of the analyzed datasets, thus providing better results.
- "Fusionstrap" combines observations from filtered datasets using the Bootstrap Rotation technique. This specific combination of techniques can help to better adapt to certain types of variables.

These results suggest that "Fusionstrap" has demonstrated a significant level of utility in generating synthetic data in the context of preserving essential features of the original data sets.

RQ1.3: To what extent does "Fusionstrap" resolve class imbalances compared to CTGAN and SYNTHPOP?

"Fusionstrap" provides a flexible and adaptable approach to handle class imbalance in synthetic datasets. By means of precise parameter setting, the methodology can be trained to pay special attention and balance the class distribution for selected variables. This functionality allowed for superior results in resolving class imbalance compared to the CTGAN and SYNTHPOP alternatives. Detailed analysis of the results (Section 6.3.3) reveals that, for the specific variables selected for imbalance correction, "Fusionstrap" was able to significantly reduce the disparities between classes, thus validating its effectiveness in handling this complex problem. This feature strengthens the position of "Fusionstrap" as a promising option for addressing class imbalance in the context of synthetic data generation.

RQ2: To what extent can "Fusionstrap" ensure the confidentiality of the original data?

RQ2.1: How to quantify the risk of disclosure of synthetic data?

To evaluate to what extent "Fusionstrap" ensures the confidentiality of the original data, we focused on the method based on data holdout and on specific evaluation metrics (section 3.5). Also, we used the disclosure risk testing method (Section 6.3.2.2). This approach allowed the quantification of the probability of re-identification of individuals in the synthetic data.

RQ2.2: How well does "Fusionstrap" protect the privacy of the original data compared to other methods (CTGAN and SYNTHPOP)?

The detailed analysis of the results on the three datasets, namely US Census, Diabetes Prediction and AIDS (Section 6.3.2), reveals that "Fusionstrap" stands out as an effective method in protecting data privacy

compared to CTGAN and Synthpop. "Fusionstrap" achieves lower values than these two methods in tests such as KS and CS on all three datasets, indicating a greater similarity between the distribution of synthetic and real data, which denotes better privacy protection. Regarding the risk of identification (re-identification accuracy), "Fusionstrap" presents in most cases lower values than the other two methods. For example, for the US Census dataset, "Fusionstrap" records the lowest re-identification accuracy (0.31) compared to Synthpop (0.67) and CTGAN (0.68). This result indicates a decrease in the risk of exposure of sensitive data in the context of using "Fusionstrap". In conclusion, the obtained results indicate that "Fusionstrap" is at the forefront in terms of data privacy protection, offering greater similarity to real data and a reduced risk of re-identification. This aspect places it in a superior position to CTGAN and Synthpop in the specific context of these analyses. Based on the DCR and NNDR tests, "Fusionstrap" showed superior values compared to CTGAN and Synthpop. These results indicate a better match of the synthetic data generated by "Fusionstrap" to the essential features of the real data sets, thus reflecting an increased quality in the generation of the synthetic data. Regarding the Logistic Detection test, "Fusionstrap" showed lower values, meaning that the generated synthetic data is less likely to be detected as artificial compared to those generated by CTGAN and Synthpop. This feature underlines the ability of "Fusionstrap" to produce synthetic data less susceptible to identification and distinction, thus strengthening its effectiveness in protecting privacy. These additional findings support the conclusions that "Fusionstrap" establishes itself as a robust option in managing the risk of exposure of sensitive data and offers significant advantages in privacy protection compared to CTGAN and Synthpop alternatives.

As a final conclusion, our research indicates that the performance of methods in the generation of synthetic data must be analyzed through the prism of the requirements of the projects and the characteristics of the data sets used. Certain characteristics of data sets can influence the results and make a method more suitable for certain scenarios. In general, there is no one-size-fits-all approach to all data sets. The experiments carried out in our research demonstrated that:

- Synthpop scores superior in terms of similarity to real distributions in most cases.
- “Fusionstrap” shows good performance in handling class imbalances and privacy risk.
- CTGAN indicated a relatively lower performance compared to the other two methods in most of the analyzed scenarios.

This evaluation highlights that "Fusionstrap" is a viable option, especially in the context of class imbalance management and privacy protection, with solid performance but room for improvement in ensuring utility for some specific variables. Finally, this research opens doors for further development of synthetic data generation methods and continuous improvement of sensitive data protection.

7.3 Limitations

The ‘Fusionstrap” method represents an innovative approach for generating synthetic data, however, as the research results highlight, the following limitations must be taken into account:

- Configuration Complexity: Optimizing the Fusionstrap parameters involved detailed expertise and fine-tuning. On the "AIDS" data set, for example, a deep understanding of the interactions between variables was required to achieve the desired results.
- Variable Performance as a function of dataset: For example, “Fusionstrap” achieved notable results on the "US Census" dataset, but performed more modestly on the "Diabetes Prediction" dataset. This aspect indicates that its performance may vary depending on the specific characteristics of the data set used.
- Vulnerability to reduced data variability: In the case of the "Diabetes Prediction" dataset, Fusionstrap had difficulty preserving the distribution of low-variance features such as "blood_glucose_level".
- Computational and time resources: Generating synthetic data can involve significant computing resources, thus limiting the widespread implementation and use of "Fusionstrap". Applying “Fusionstrap” to large datasets such as the US Census imposed a significant demand on computational resources, which can affect efficiency and execution time.

7.4 Future work

This section is dedicated to further development directions of the “Fusionstrap” framework, given the complexity and importance of handling class imbalances and ensuring data privacy. Several research directions can be considered to improve and extend this framework:

- Expanding Data Generation Capability: Further research may aim to optimize the architecture, parameters and associated technologies to increase the efficiency and accuracy of the method. An important consideration for further development of “Fusionstrap” is expanding its ability to generate synthetic data for a wider range of data types and domains. This may involve improving how “Fusionstrap” handles textual data, images, or other domain-specific data formats.
- Refined Sampling Strategies: Continued research to improve the subsampling strategies used in “Fusionstrap”, such as the Circular Nearest Neighbor (CNN) algorithm, to ensure a more accurate selection of synthetic data. Evaluation and optimization of these strategies can lead to higher quality synthetic data.
- Development of Advanced Evaluation Technologies: Creation of more advanced and rigorous evaluation methods to measure the quality of synthetic data generated by “Fusionstrap”. This may involve developing custom metrics or using more advanced privacy risk assessment technologies, such as MIA (Member Inference Attack).
- Domain-Specific Customization: Adapting “Fusionstrap” to the specific requirements of different domains or industries to effectively address their specific class and privacy imbalance issues.
- Integration with Machine Learning Technologies: Integration of “Fusionstrap” with advanced machine learning technologies such as deep learning or transfer learning to fully leverage the generated synthetic data.
- Optimization of Computational Resources and Time: Continued efforts to optimize the use of computational resources and time required to run “Fusionstrap”, especially for large datasets, to make the framework more accessible and resource-efficient, and of time.

These development directions could help strengthen and improve “Fusionstrap” into a powerful and versatile tool for managing class imbalances and protecting data privacy in data analytics and machine learning, with the potential to address a broad spectrum of problems and applications.

Bibliography

[1] Oluoch, T., Santas, X., Kwaro, D., Were, M., Biondich, P., Bailey, C., ... & de Keizer, N. (2012). The effect of electronic medical record-based clinical decision support on HIV care in resource-constrained settings: a systematic review. *International journal of medical informatics*, *81*(10), e83-e92.

[2] Converse, S. J., Moore, C. T., & Armstrong, D. P. (2013). Demographics of reintroduced populations: estimation, modeling, and decision analysis. *The Journal of Wildlife Management*, *77*(6), 1081-1093.

[3] Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem. *Int. J. Advance Soft Compu. Appl*, *5*(3), 176-204.

[4] Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks*, *21*(2-3), 427-436.

[5] ARGYLE, L. P., & Barber, M. (2022). Misclassification and bias in predictions of individual ethnicity from administrative records. *American Political Science Review*, 1-9.

[6] Rahman, M. M., & Davis, D. N. (2013). Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, *3*(2), 224.

[7] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in neural information processing systems*, *32*.

[8] Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke creation of synthetic data in R. *Journal of statistical software*, *74*, 1-26.

[9] Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective*. Routledge.

[10] Booth, J. G., & Hall, P. (1994). Monte Carlo approximation and the iterated bootstrap. *Biometrika*, 81(2), 331-340.

[11] Drechsler, J., & Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, *105*(492), 1347-1357.

[12] Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press

[13] Pandey, P., & Pandey, M. M. (2021). *Research methodology tools and techniques*. Bridge Center

[14] Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics* (pp. 569-593). Springer, New York, NY

[15] El Emam, K., Mosquera, L., & Hoptroff, R. (2020). *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media.url:https: //learning.oreilly.com/library/view/practical-synthetic- data/ 9781492072737

[16] Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, *60*(3), 224-232.

[17] Garousi, V., Felderer, M., & Mäntylä, M. V. (2019). Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and software technology*, *106*, 101-121.

[18] Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, *51*(1), 7-15.

[19] Wang, S., Minku, L. L., & Yao, X. (2013, April). A learning framework for online class imbalance learning. In *2013 IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL)* (pp. 36-45). IEEE.

[20] Maalouf, M., & Trafalis, T. B. (2011). Rare events and imbalanced datasets: an overview. *International Journal of Data Mining, Modelling and Management*, *3*(4), 375-388.

[21] Oommen, T., Baise, L. G., & Vogel, R. M. (2011). Sampling bias and class imbalance in maximum-likelihood logistic regression. *Mathematical Geosciences*, *43*, 99-120.

[22] Brodley, C. E., & Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of artificial intelligence research*, *11*, 131-167.

[23] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

[24] He, H., & Ma, Y. (2013). Imbalanced Learning-Foundations, Algorithms, and Applications, New Jersey: The Institute of Electrical and Electronics Engineers.
[25] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, *6*(5), 429-449.
[26] Elkan, C. (2001, August). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence* (Vol. 17, No. 1, pp. 973-978). Lawrence Erlbaum Associates Ltd.
[27] Cao, P., Zhao, D., & Zaiane, O. (2013, April). An optimized cost-sensitive SVM for imbalanced data learning. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 280-292). Berlin, Heidelberg: Springer Berlin Heidelberg.
[28] Shen, F., Wang, R., & Shen, Y. (2020). A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach. *Technological and Economic Development of Economy*, *26*(2), 405-429.
[29] Hasanin, T., & Khoshgoftaar, T. (2018, July). The effects of random undersampling with simulated class imbalance for big data. In *2018 IEEE international conference on information reuse and integration (IRI)* (pp. 70-79). IEEE.
[30] Elhassan, T., & Aljurf, M. (2016). Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method. *Global J Technol Optim S*, *1*, 2016.
[31] Choirunnisa, S., & Lianto, J. (2018, November). Hybrid method of undersampling and oversampling for handling imbalanced data. In *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 276-280). IEEE.
[32] Nigam, A. K., & Rao, J. N. K. (1996). On balanced bootstrap for stratified multistage samples. *Statistica Sinica*, 199-214.
[33] Brownlee, J. (2020). Random oversampling and undersampling for imbalanced classification. *Machine learning mastery*.
[34] Kunz, P. J., & Zoubir, A. M. (2023, June). Heterogeneity-Stratified Bootstrap Oversampling for Training a Spoiled Food Detector. In *2023 24th International Conference on Digital Signal Processing (DSP)* (pp. 1-5). IEEE.
[35] Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics* (pp. 569-593). Springer, New York, NY.
[36] Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (No. 1). Cambridge university press.
[37] Hall, P., & Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 757-762.
[38] Chernick, M. R. (2011). *Bootstrap methods: A guide for practitioners and researchers*. John Wiley & Sons.
[39] Brownlee, Jason. "A Gentle Introduction to the Bootstrap Method". *Machine Learning Mastery, May 25th,2018.* https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/.
[40] Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. *Monographs on statistics and applied probability*, *57*, 1-456. London, New York: Chapman and Hall.
[41] Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. The Annals of Statistics, 14(4), 1261-1350.
[42] Mooney, C. Z., Duval, R. D., & Duvall, R. (1993). *Bootstrapping: A nonparametric approach to statistical inference* (No. 95). sage.
[43] Shao, J., & Tu, D. (1995). The Jackknife and Bootstrap. Springer-Verlag.
[44] https://online.stat.psu.edu/stat555/node/119/
[45] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, *78*(382), 316-331.
[46] https://www.investopedia.com/terms/s/statistics.asp
[47] https://www.statista.com/statistics-glossary/definition/398/distribution_function/
[48] Efron, B. (1979). Bootstrap methods: another look at the jackknife annals of statistics 7: 1–26. *View Article PubMed/NCBI Google Scholar*, *24*.
[49] https://www.investopedia.com/terms/c/confidenceinterval.asp
[50] DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical science*, *11*(3), 189-228.

[51] Orloff, J., & Bloom, J. (2014). Bootstrap confidence intervals. *Retrieved from MIT OpenCourseWare website: https://ocw. mit. edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading24. pdf*.
[52] https://statisticsbyjim.com/basics/parameter-vs-statistic/
[53] Forst, Jim. "Introduction to Bootstrapping in Statistics with an Example". *Statistics by Jim.* https://statisticsbyjim.com/hypothesis-testing/bootstrapping/.
[54] Zoubir, A. M., & Iskandler, D. R. (2007). Bootstrap methods and applications. *IEEE Signal Processing Magazine*, *24*(4), 10-19.
[55] Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, *87*(419), 738-754.
[56] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.
[57] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). *Modeling tabular data using conditional gan. Advances in neural information processing systems*, 32.
[58] Moon, J., Jung, S., Park, S., & Hwang, E. (2020). Conditional Tabular GAN-Based Two-Stage Data Generation Scheme for Short-Term Load Forecasting. IEEE Access, 8, 205327–205339. https://doi.org/10.1109/access.2020.3037063
[59] Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke creation of synthetic data in R. *Journal of statistical software*, *74*, 1-26.
[60] Raab, G. M., & Nowok, B. (2017). Inference from fitted models in Synthpop. *R Vignette, URL https://cran. r-project. org/web/packages/Synthpop/vignettes/inference. pdf*.
[61] El Emam, K., Mosquera, L., Fang, X., & El-Hussuna, A. (2022). Utility metrics for evaluating synthetic health data generation methods: validation study. *JMIR medical informatics*, *10*(4), e35734.
[62] Sedgwick, P. (2012). Pearson's correlation coefficient. *Bmj*, *345*.
[63] Jordon, J., Yoon, J., & Van Der Schaar, M. (2018, September). PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*.
[64] Agresti, A. (2012). *Categorical data analysis* (Vol. 792). John Wiley & Sons.
[65] Platzer, M., & Reutterer, T. (2021). Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in big Data*, *4*, 679939.
[66] Platzer, M., & Reutterer, T. (2021). Holdout-based empirical assessment of mixed-type synthetic data. Frontiers in big Data, 4, 679939. https://www.frontiersin.org/articles/10.3389/fdata.2021.679939/full
[67] Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.
[68] Chakravart, L. Roy (1967). Handbook of Methods of Applied Statistics, vol. I.
[69] Kirkman, T.W. (1996) Statistics to Use: Kolmogorov-Smirnov test. (Accessed 10 Feb 2010)
[70] Wang, Z., Myles, P., & Tucker, A. (2019, June). Generating and evaluating synthetic UK primary care data: preserving data utility & patient privacy. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 126-131). IEEE.
[71] Ugoni, A., & Walker, B. F. (1995). The Chi square test: an introduction. *COMSIG review*, *4*(3), 61.
[72] Pandis, N. (2016). The chi-square test. *American journal of orthodontics and dentofacial orthopedics*, *150*(5), 898-899.
[73] Sharpe, D. (2015). Chi-square test is statistically significant: Now what?. *Practical Assessment, Research, and Evaluation*, *20*(1), 8.
[74] Hernadez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2023). Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. *Methods of Information in Medicine*.
[75] Abay, N. C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B., & Sweeney, L. (2019). Privacy preserving synthetic data release using deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18* (pp. 510-526). Springer International Publishing.
[76] LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395-2399.
[77] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Belmont, CA: Wadsworth. *International Group*, *432*(151-166), 9.
[78] https://www.javatpoint.com/logistic-regression-in-machine-learning

[79] Kyurkchiev, N., & Markov, S. (2015). Sigmoid functions: some approximation and modelling aspects. *LAP LAMBERT Academic Publishing, Saarbrucken*, *4*.
[80] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.
[81] Walsh, A. (1987). Teaching understanding and interpretation of logit regression. *Teaching sociology*, 178-183.
[82] Brzezinski, J. R., & Knafl, G. J. (1999, September). Logistic regression modeling for context-based classification. In *Proceedings. Tenth International Workshop on Database and Expert Systems Applications. DEXA 99* (pp. 755-759). IEEE.
[83] Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.
[84] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.
[85] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.
[86] Onesmus, M. (2020). *Introduction to Random Forest in Machine Learning* [png]. Section. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/
[87] Assefa, S. A., Dervovic, D., Mahfouz, M., Tillman, R. E., Reddy, P., & Veloso, M. (2020, October). Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance* (pp. 1-8).
[88] Brenninkmeijer, B., de Vries, A., Marchiori, E., & Hille, Y. (2019). On the generation and evaluation of tabular data using GANs. *PhD diss., Radboud University*.
[89] Mendelevitch, O., & Lesh, M. D. (2021). Fidelity and privacy of synthetic medical data. *arXiv preprint arXiv:2101.08658*.
[90] Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.
[91] Gussenbauer, J., Kowarik, A., Kalcher, K., & Platzer, M. (2021). AI-Based Privacy Preserving Census (like) Data Publication.
[92] Platzer, M., & Reutterer, T. (2021). Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in big Data*, *4*, 679939.
[93] Mendes Júnior, P. R., De Souza, R. M., Werneck, R. D. O., Stein, B. V., Pazinato, D. V., de Almeida, W. R., ... & Rocha, A. (2017). Nearest neighbors distance ratio open-set classifier. *Machine Learning*, *106*(3), 359-386.
[94] Guillaudeux, M., Rousseau, O., Petot, J., Bennis, Z., Dein, C. A., Goronflot, T., ... & Gourraud, P. A. (2023). Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *NPJ Digital Medicine*, *6*(1), 37.
[95] Zhao, Z., Kunar, A., Birke, R., & Chen, L. Y. (2021, November). Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning* (pp. 97-112). PMLR.
[96] Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *Annals of translational medicine*, *4*(1).
[97] Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, *173*(3), 781-800.
[98] Bekkar, M., & Alitouche, T. A. (2013). Imbalanced data learning approaches review. *International Journal of Data Mining & Knowledge Management Process*, *3*(4), 15.
[99] Benali, F., Bodénès, D., Labroche, N., & de Runz, C. (2021). MTCopula: Synthetic complex data generation using copula. In *23rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP)* (pp. 51-60).
[100] Kannan, K. S., Manoj, K., & Arumugam, S. (2015). Labeling methods for identifying outliers. *International Journal of Statistics and Systems*, *10*(2), 231-238.
[101] Hall, P., & Horowitz, J. L. (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica: Journal of the Econometric Society*, 891-916.
[102] Kohavi, R., & Becker, B. (1994). Data mining and visualization. Silicon graphics. *Extraction from the*.
[103] Westenberg, A. A. (1965). Applied Physics Laboratory The Johns Hopkins University Silver Spring, Maryland. *MIDWEST RESEARCH INSTITUTE*, 19.
[104] https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[105] Hammer, S. M., Squires, K. E., Hughes, M. D., Grimes, J. M., Demeter, L. M., Currier, J. S., ... & Cook, J. C. (1997). A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine*, *337*(11), 725-733.
[106] Pagès, J. (2004). Analyse factorielle de donnees mixtes: principe et exemple d'application. *Revue de statistique appliquée*, *52*(4), 93-111.

Appendices

A. Description of data sets

The relevance of the data sets selected for the experimental part also derives from the nature of the variables that compose them. For a better understanding of the experimental objectives and results, it is necessary to describe the variables.

Table A.1: US census variables

Variables	Explanation and relevance	Variable type
Age	Age in years. Integer greater than 0 (zero);	Numerical
Work class	General term for the employment status of the person (e.g., Private, Selfempnotinc, Selfempinc, Federalgov, Localgov, Stategov, Withoutpay, Neverworked)	Categorical
fnlwgt	Final weight - the number of people who meet the same characteristics (attributes) considered census inputs - integer greater than 0	Numerical
Education	The highest level of education attained by an individual (e.g., Bachelors, Somecollege, 11 th, HSgrad, Profschool, Assocacdm, Assocvoc, 9 th, 7th8th, 12 th, Masters, 1st4th, 10 th, PhDs, 5th6th, Preschool.	Categorical
Marital status of a person	e.g. married, divorced, never married, separated, widowed, etc.;	Categorical
Occupation	The general type of occupation of an individual (Techsupport, Craftrepair, Otherservice, Sales, Execmanagerial, Profspecialty, Handlerscleaners, Machineopinspct, Admclerical, Farmingfishing, Transportmoving, Privhouseserv, Protectiveserv, ArmedForces.	Numerical Each type was labeled with a whole number greater than 0, from 1 to 14.
Relationship	What this individual is in relation to others. For example, an individual could be a spouse, child, relative, unmarried, etc. Each entry has a single relationship attribute and is somewhat redundant with the marital status, so this feature will not be used in the analysis	Categorical
Race	Race description of an individual (White, AsianPacIslander, AmerIndianEskimo, Other, Black	Categorical
Sex	The biological sex of the individual.	Categorical
Capital-gain	Capital gains for an individual: integer greater than or equal to 0	Numerical
Capital-loss	Capital loss for an individual: integer greater than or equal to 0	Numerical
Hours-per-week	The hours an individual has reported to work per week: continuous.	Numerical
Native-country	Country of origin for an individual	Categorical
The label	Whether or not an individual makes more than $50,000 annually (<=50k, >50k).	Categorical

Table A.2: Diabet Prediction variables

Variables	Explanation and relevance	Variable type
Gender	It involves classifying individuals as male or female. This demographic factor can influence various aspects of health.	Categorical (0 or 1)
Age	The individual's age provides insight into the potential health risks associated with certain age groups and susceptibility to certain diseases.	Numerical
Hypertension	High blood pressure is a medical condition characterized by persistently elevated blood pressure in the arteries with significant health risk if left unmanaged.	Categorical (0 or 1)
Heart disease	It is a broad term that encompasses various cardiovascular disorders that can affect the overall functioning of the heart.	Categorical (0 or 1)
Smoking-history	Smoking history indicates whether a person has a past or present habit of smoking tobacco products. Smoking is a well-known risk factor for many health problems.	Categorical
bmi	Body mass index (BMI) provides an estimate of whether a person's weight is in a healthy range, or whether they are underweight or obese. BMI is commonly used as a screening tool to assess the risk of weight-related health problems.	Numerical
HbA1c_level	The HbA1c (hemoglobin A1c) level is a laboratory test that measures the average level of sugar (glucose) in the blood over the last 2-3 months. It is commonly used in the diagnosis and management of diabetes mellitus.	Numerical
Blood_glucose_level	Blood glucose level refers to the concentration of glucose (sugar) in the blood. Abnormal blood glucose levels, either too high (hyperglycemia) or too low (hypoglycemia), can be associated with various health conditions, especially diabetes.	Numerical
Diabetes	Diabetes is a chronic condition characterized by high blood sugar levels due to insulin deficiency or ineffective use of insulin. It requires careful management and ongoing monitoring to prevent complications, impacting overall health and requiring lifestyle changes, treatments and regular medical care.	Categorical (0 or 1)

Table A.3: AIDS variables

Variables	Explanation and relevance	Variable type
Age	Age in years. Integer greater than 0 (zero);	Numerical
wtkg	It contains information about the person's weight in kilograms and can be used to analyze relationships between weight and other characteristics or variables in the dataset. Usually, in medical or health analysis, body weight can be an important indicator for assessing a person's health status and identifying risk factors.	Numerical
hemo	The level of hemoglobin in the blood can provide information about the health of the circulatory system and can be used to assess the oxygenation status of tissues and potential problems related to anemia or other conditions.	Numerical (0 or 1)
homo	Indicates whether the individual is heterosexual (0) or homosexual (1). In HIV/AIDS studies, analysis of sexual orientation can help to understand the risk factors and specific needs of different groups. At the same time, consideration must be given to protecting the privacy of sensitive data, such as sexual orientation, to ensure respect for individual rights and research ethics.	Numerical (0 or 1)
drugs	It indicates whether or not an individual uses drugs in the context of the AIDS-related condition. This variable could be used to analyze the impact of drug use on the risk or progression of AIDS-related disease within the data set.	Numerical (0 or 1)
zprior	Binary indicator that provides contextual information about a patient's previous status or history.	Numerical (0 or 1)
preanti	Reflect a pre-existing condition or previous intervention in the treatment or management of HIV/AIDS.	Numerical
strat	Categorical or grouping variable, possibly related to the stratification of patients into different categories.	Numerical
symptom	Indicates the presence or absence of certain symptoms related to HIV/AIDS.	Numerical (0 or 1)
treat	Shows the type or status of treatment received by the patient.	Numerical (0 or 1)
offtrt	Means if a patient has exited or discontinued treatment.	Numerical (0 or 1)
cd40, cd420, cd496, cd80, cd820	Represent biological markers or measured values associated with laboratory tests for HIV/AIDS patients.	Numerical
cens	Reflects whether a patient was censored or not in the study.	Numerical (0 or 1)
days	The number of days or the specific period associated with certain events or measurements.	Numerical
arms	Reflect the groups or categories into which the patients in the study are divided.	Numerical

B. Utility evaluation

B1. Results for the Hellinger distance

This appendix provides the detailed results of the utility assessment using the Hellinger distance in our study. The Hellinger distance served as a metric for quantifying the similarity of probability distributions between authentic and synthetic data. We introduce this appendix to provide a detailed insight into how Fusionstrap compares to alternative methods, highlighting the results obtained for specific variables and the areas where the synthetic data best reflect the real data or show significant differences.

Table B.1.1: Hellinger distance for US census

Hellinger distance	"Fusionstrap"	CTGAN	SYNTHPOP
age	0.001503	0.019314	0.000514
workclass	0.078416	0.017729	0.000113
fnlwgt	0.008189	0.006950	0.042397
education	0.043857	0.011028	0.000489
education.num	0.087514	0.637059	0.000489
marital.status	0.076949	0.015578	0.000055
occupation	0.039127	0.030049	0.000389
relationship	0.007804	0.010298	0.000341
race	0.018131	0.002459	0.000150
sex	0.000057	0.004805	0.000120
capital gain	0.589907	0.646915	0.000338
capital loss	0.091504	0.663078	0.000435
hours-per-week	0.311244	0.599965	0.000547
native.country	0.137408	0.030104	0.000321
income	0.048115	0.002386	0.000050

Table B.1.2: Hellinger distance for Diabet Prediction

Hellinger distance	"Fusionstrap"	CTGAN	SYNTHPOP
gender	0.015110	0.000503	0.000004
age	0.109009	0.390010	0.001567
hypertension	0.003441	0.015873	0.000001
heart_disease	0.176758	0.025038	0.000001
smoking_history	0.024159	0.025451	0.000063
bmi	0.174631	0.166406	0.008958
HbA1c_level	0.064802	0.221252	0.000425
blood_glucose_level	0.173604	0.097442	0.000498
diabetes	0.172754	0.006338	0.000091

Table B.1.3: Hellinger distance for AIDS

Hellinger distance	"Fusionstrap"	CTGAN	SYNTHPOP
age	0.006968	0.002495	1.200386e-03
wtkg	0.104049	0.220726	2.780126e-03
hemo	0.010972	0.000234	1.390168e-04
homo	0.003339	0.015234	2.901719e-05
drugs	0.009573	0.024315	7.721386e-04
karnof	0.024923	0.001064	3.739790e-04
oprior	0.021312	0.000231	1.611478e-04
z30	0.000389	0.000529	2.615506e-04
zprior	0.000000	0.000000	0.000000e+00
preanti	0.247569	0.066064	6.758973e-03
race	0.002058	0.000221	1.837184e-04
gender	0.009469	0.000004	0.000000e+00
str2	0.000173	0.002511	3.202560e-04
strat	0.001292	0.013438	7.377601e-04
symptom	0.008678	0.001305	1.743516e-05
treat	0.006225	0.001702	8.281419e-07
offtrt	0.011319	0.001832	1.138289e-04
cd40	0.009836	0.010994	2.448701e-03
cd420	0.007519	0.016215	3.046304e-03
cd496	0.294832	0.207047	1.996838e-01
r	0.002971	0.001585	2.802622e-05
cd80	0.008362	0.008641	6.968380e-03
cd820	0.007117	0.005718	5.711251e-03
cens	0.006493	0.000870	2.098416e-07
days	0.020985	0.011047	2.912420e-03
arms	0.000188	0.015396	8.009934e-04

B2. Difference synthetic-original between the correlation coefficients

Table B.2.1: Synthetic-original between the correlation coefficients for US census

Variables	"Fusionstrap"	CTGAN	SYNTHPOP
age+capital.gain	0.030877356	0.037375304	0.005221219
age+capital.loss	0.026183272	0.035784823	0.003031034
age+education	0.182766844	0.150876996	0.004409319
age+education.num	0.010447421	0.000537373	0.006960134
age+hours.per.week	0.034251491	0.023278500	0.005711743
age+marital.status	0.553763463	0.463126863	0.005583640
age+native.country	0.046138620	0.015898270	0.029275029
age+occupation	0.158382292	0.115456744	0.021877432
age+race	0.026661355	0.035372352	0.008943997
age+relationship	0.460738150	0.354427556	0.008086746
age+sex	0.086566972	0.023515518	0.006441464
age+workclass	0.188233334	0.126756490	0.007662513

age+income	0.207938504	0.094016119	0.013395730
capital.gain+capital.loss	0.093977583	0.021983136	0.003723434
capital.gain+education	0.105380997	0.174281754	0.019193401
capital.gain+education.num	0.076850129	0.060448725	0.011419759
capital.gain+hours.per.week	0.013959503	0.033641239	0.014702604
capital.gain+marital.status	0.174703668	0.064196263	0.010184496
capital.gain+native.country	0.174821315	0.001132947	0.008573591
capital.gain+occupation	0.101131931	0.081133429	0.020793276
capital.gain+race	0.112191410	0.008219819	0.003775339
capital.gain+relationship	0.199006012	0.076675292	0.013556822
capital.gain+sex	0.098862061	0.045832965	0.015703870
capital.gain+workclass	0.130617382	0.082346235	0.011939874
capital.gain+income	0.262891337	0.217380817	0.000781094
capital.loss+education	0.094303478	0.068067926	0.007805528
capital.loss+education.num	0.052414994	0.032128934	0.004329624
capital.loss+hours.per.week	0.005309023	0.029478504	0.005890569
capital.loss+marital.status	0.085814431	0.066350669	0.001239989
capital.loss+native.country	0.086189957	0.018673666	0.009815501
capital.loss+occupation	0.058623239	0.058660383	0.012507548
capital.loss+race	0.056997577	0.011784708	0.016085378
capital.loss+relationship	0.107487160	0.074136948	0.003439648
capital.loss+sex	0.040345461	0.041952819	0.000587995
capital.loss+workclass	0.117120128	0.026270330	0.020618690
capital.loss+income	0.167554604	0.148626271	0.021713326
education+education.num	0.634494573	0.960033808	0.000037300
education+hours.per.week	0.067439118	0.168053433	0.006971713
education+marital.status	0.041269957	0.033399522	0.010239546
education+native.country	0.048304919	0.079586349	0.003988406
education+occupation	0.096610667	0.131463247	0.001843328
education+race	0.010890195	0.017821044	0.010660941
education+relationship	0.050639450	0.044513774	0.011122570
education+sex	0.094479051	0.002300630	0.020364687
education+workclass	0.025329307	0.037538217	0.007809198
education+income	0.257007073	0.177263663	0.027675521
education.num+hours.per.week	0.007984622	0.072926457	0.000075432
education.num+marital.status	0.184574930	0.086573435	0.023848846
education.num+native.country	0.034203186	0.233640610	0.020001283
education.num+occupation	0.248485787	0.444863032	0.006190355
education.num+race	0.040873661	0.082253759	0.020755921
education.num+relationship	0.194417328	0.121974978	0.026614011
education.num+sex	0.161131251	0.027660482	0.006361728
education.num+workclass	0.092743399	0.154426235	0.022809229
education.num+income	0.242089953	0.279157835	0.029387507
hours.per.week+marital.status	0.140313243	0.203863940	0.021882481
hours.per.week+native.country	0.064882361	0.011324728	0.000457134
hours.per.week+occupation	0.165818904	0.233054324	0.029611945

hours.per.week+race	0.006926266	0.030907664	0.001238525
hours.per.week+relationship	0.175960845	0.275153605	0.004386715
hours.per.week+sex	0.178773370	0.210192324	0.014663101
hours.per.week+workclass	0.068054866	0.136123964	0.035740200
hours.per.week+income	0.018204140	0.201909642	0.039659307
marital.status+native.country	0.037399007	0.021750727	0.008823263
marital.status+occupation	0.032238306	0.057629448	0.006356311
marital.status+race	0.009418197	0.011884679	0.006820297
marital.status+relationship	0.347863964	0.387149201	0.001712030
marital.status+sex	0.300718944	0.289271122	0.009874321
marital.status+workclass	0.029477802	0.021036201	0.002469485
marital.status+income	0.071807155	0.225012527	0.003591563
native.country+occupation	0.008509943	0.020069484	0.018063605
native.country+race	0.330574080	0.361924321	0.001528088
native.country+relationship	0.044334135	0.019225623	0.013207778
native.country+sex	0.057915048	0.058260766	0.017776662
native.country+workclass	0.062546889	0.041884506	0.011663536
native.country+income	0.335342372	0.025824777	0.031699724
occupation+race	0.021447279	0.016434843	0.014843947
occupation+relationship	0.056275525	0.091738431	0.009699685
occupation+sex	0.295910698	0.276802644	0.010081525
occupation+workclass	0.111331513	0.127131257	0.004955766
occupation+income	0.105343126	0.138732837	0.033818699
race+relationship	0.012216267	0.022424179	0.004334013
race+sex	0.034964953	0.012805125	0.021810190
race+workclass	0.015452050	0.015019757	0.019363881
race+income	0.172801801	0.028620832	0.023061911
relationship+sex	0.461939166	0.431057022	0.001899806
relationship+workclass	0.045822576	0.020119558	0.010206591
relationship+income	0.154089283	0.165483406	0.000728824
sex+workclass	0.007233333	0.026221975	0.029312744
sex+income	0.087239686	0.016599846	0.009359481
workclass+income	0.324506503	0.024203669	0.031039602
MIN	**0.005309023**	**0.000537373**	**0.000037300**
MAX	**0.634494573**	**0.960033808**	**0.039659307**
PERCENT MINIMUM VALUES	**7**	**10**	**87**

	the greatest value
	the lowest value

Table B.2.2: Synthetic-original between the correlation coefficients for Diabetes Prediction

Variables	"Fusionstrap"	CTGAN	SYNTHPOP
age+HbA1c_level	0.182975004	0.095621293	0.013621623
age+blood_glucose_level	0.18742279	0.066419895	0.029573192

age+bmi	0.047404811	0.041960622	0.011949234
age+gender	0.242200629	0.061252577	0.031321632
age+heart_disease	0.006595069	0.076150249	0.001031305
age+hypertension	0.001914897	0.05509954	0.007515468
age+smoking_history	0.157416932	0.160732263	0.005594546
age+diabetes	0.351625979	0.064810368	0.090552499
HbA1c_level+blood_glucose_level	0.235937311	0.03995681	0.004069638
HbA1c_level+bmi	0.107994331	0.019540153	0.020937886
HbA1c_level+gender	0.335330451	0.001740235	0.014096587
HbA1c_level+heart_disease	0.279147575	0.072340508	0.02947701
HbA1c_level+hypertension	0.260512655	0.029757724	0.034477795
HbA1c_level+smoking_history	0.107124803	0.176140394	0.045412704
HbA1c_level+diabetes	0.394827733	0.047012722	0.029523224
blood_glucose_level+bmi	0.110707561	0.038980076	0.000760228
blood_glucose_level+gender	0.364879923	0.047021006	0.003314622
blood_glucose_level+heart_disease	0.284140289	0.012375275	0.03234896
blood_glucose_level+hypertension	0.265012191	0.064639254	0.016345222
blood_glucose_level+smoking_history	0.137181095	0.224359351	0.018587113
blood_glucose_level+diabetes	0.413378282	0.171625083	0.029709135
bmi+gender	0.14744782	0.109771348	0.014362579
bmi+heart_disease	0.107729859	0.033703037	0.015372102
bmi+hypertension	0.035859086	0.054317907	0.007880302
bmi+smoking_history	0.108040874	0.029556079	0.008721758
bmi+diabetes	0.185527511	0.021257398	0.087136853
gender+heart_disease	0.098724616	0.000963539	0.019396628
gender+hypertension	0.17293357	0.057192331	0.010025544
gender+smoking_history	0.004700938	0.020382604	0.052038929
gender+diabetes	0.44171699	0.050641666	0.019294253
heart_disease+hypertension	0.066648308	0.067060149	0.042420228
heart_disease+smoking_history	0.010536325	0.195893776	0.063317641
heart_disease+diabetes	0.236444662	0.171414196	0.085740042
hypertension+smoking_history	0.004789477	0.085751908	0.087866908
hypertension+diabetes	0.21275362	0.183097886	0.027651244
smoking_history+diabetes	0.108190884	0.250465857	0.086943738
MIN	**0.001914897**	**0.000963539**	**0.000760228**
MAX	**0.441716990**	**0.250465857**	**0.090552499**
PERCENT MINIMUM VALUES	**11**	**19**	**67**

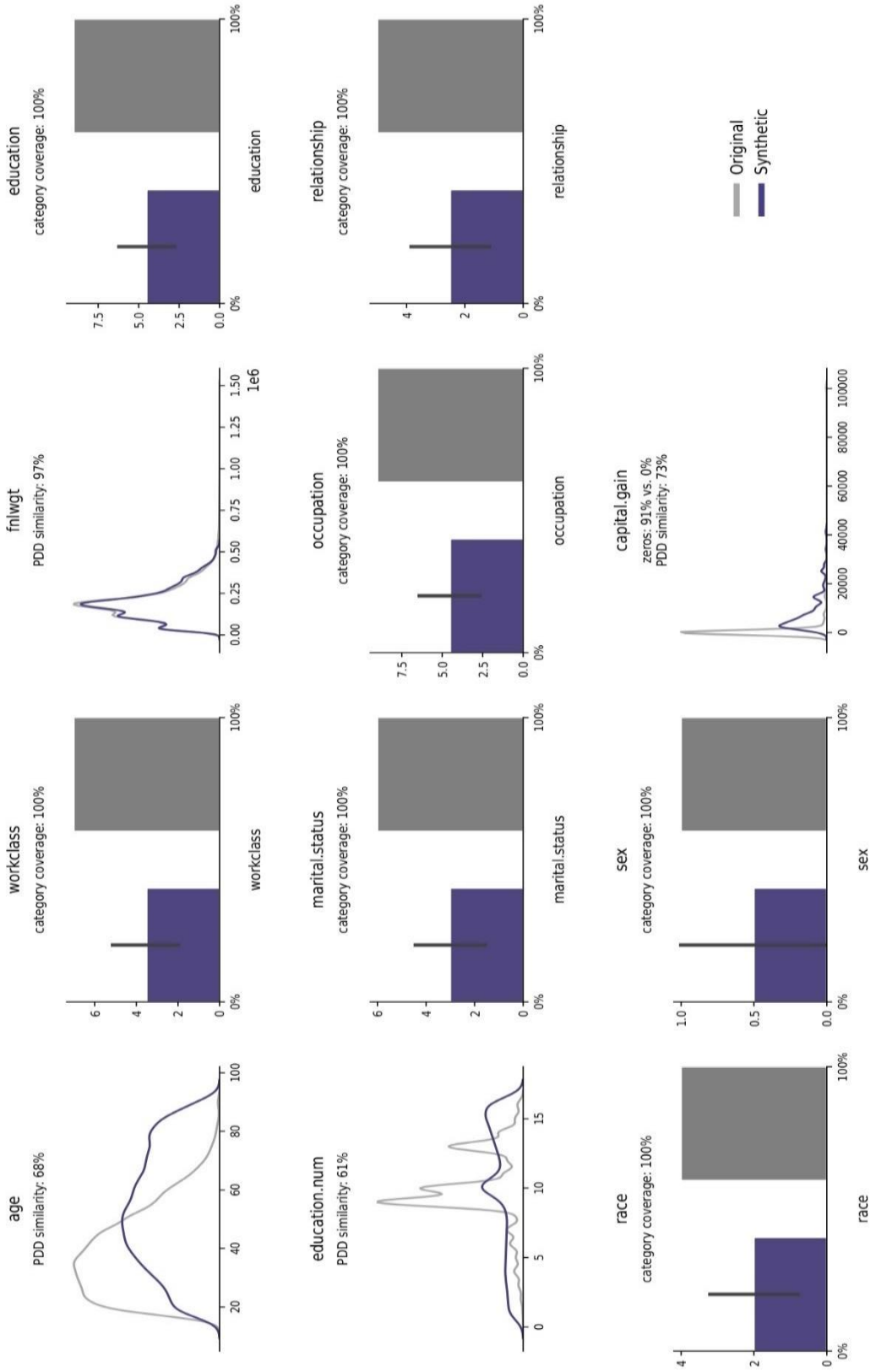
	the greatest value
	the lowest value

C. Univariate Distributions

Another way to check the usefulness of the generated data is to compare their univariate distributions with the original. Univariate distributions refer to the probability distributions of a single variable or attribute in a data set. These distributions describe how the values of that variable are spread across the data set. If the univariate distributions of the synthetic data closely resemble those of the original data, it implies that the synthetic data has successfully retained the statistical characteristics of the authentic dataset. This is a measure of utility, as it shows that the synthetic data can be used for statistical analysis without significant harm to the results.

To keep the paper to a reasonable size, we restricted the analysis of univariate distributions to the US Census data set. Thus, **Figures C.1**, **C.2,** and **C.3** capture the resemblance of the univariate distributions in the synthetic datasets produced by the three methods ("Fusionstrap", CTGAN, and SYNTHPOP) in contrast to the initial US Census dataset. Probabilistic Density Distribution (PDD) Similarity serves as an indicator for gauging the likeness between the probability density distributions of either two datasets or two variables. This measure is used in data analysis and evaluation of synthetic data to determine how well the probability density distribution (such as a normal distribution or any other distribution) of the synthetic data matches that of the real or original data. "Category coverage" refers to the extent to which the synthetic data covers all possible categories or values of a variable or attribute in the original data. Values close to 100% for these measures show that the synthetic data follow the distribution characteristics of the real data. Analyzing the values of PDD Similarity and "Category coverage" for the three approaches (Figure C.1, Figure C.2 and Figure C.3), we observe that the synthetic data effectively preserved the distribution characteristics of the real data.

Figure C.1
Univariate distributions with "Fusionstrap"



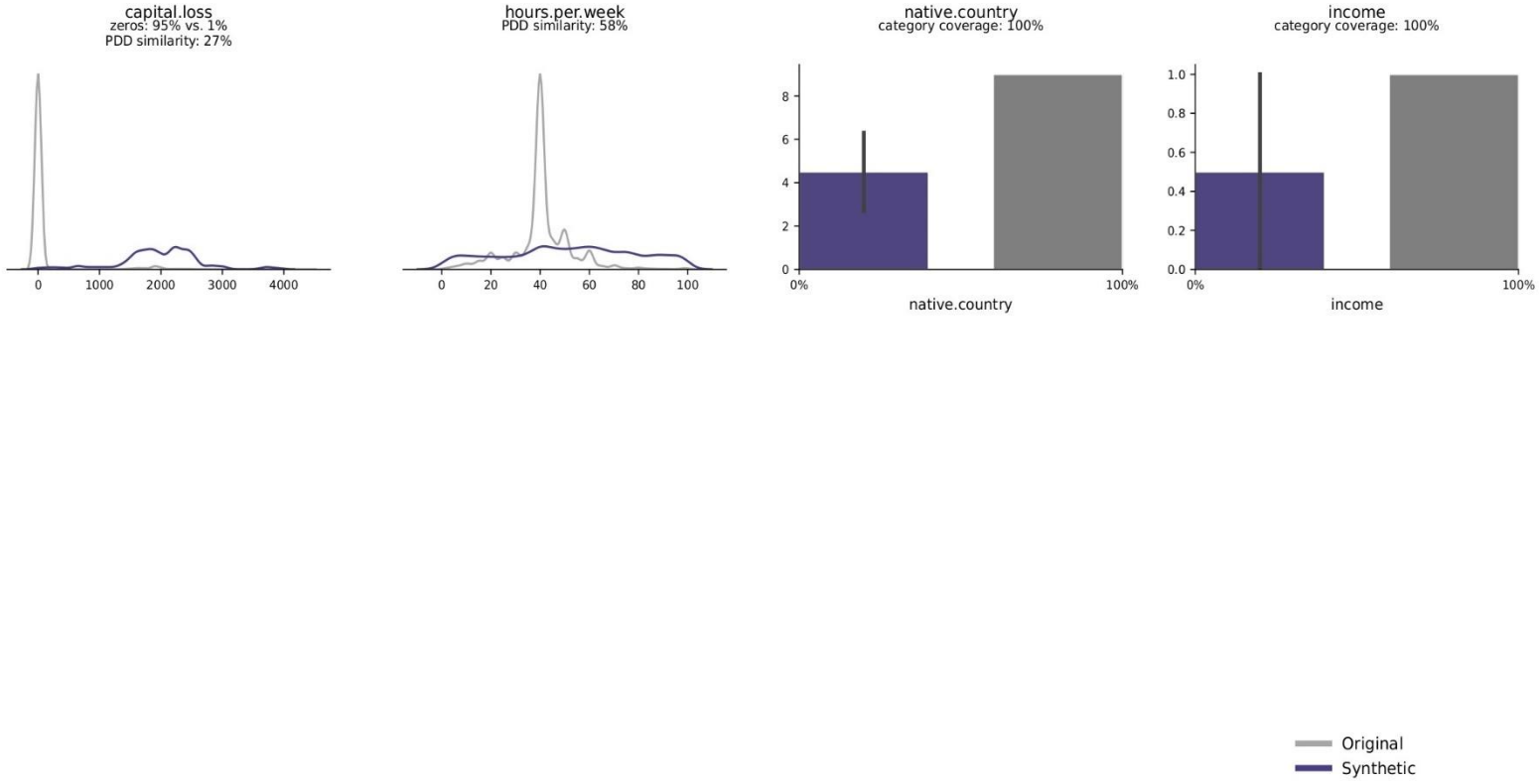
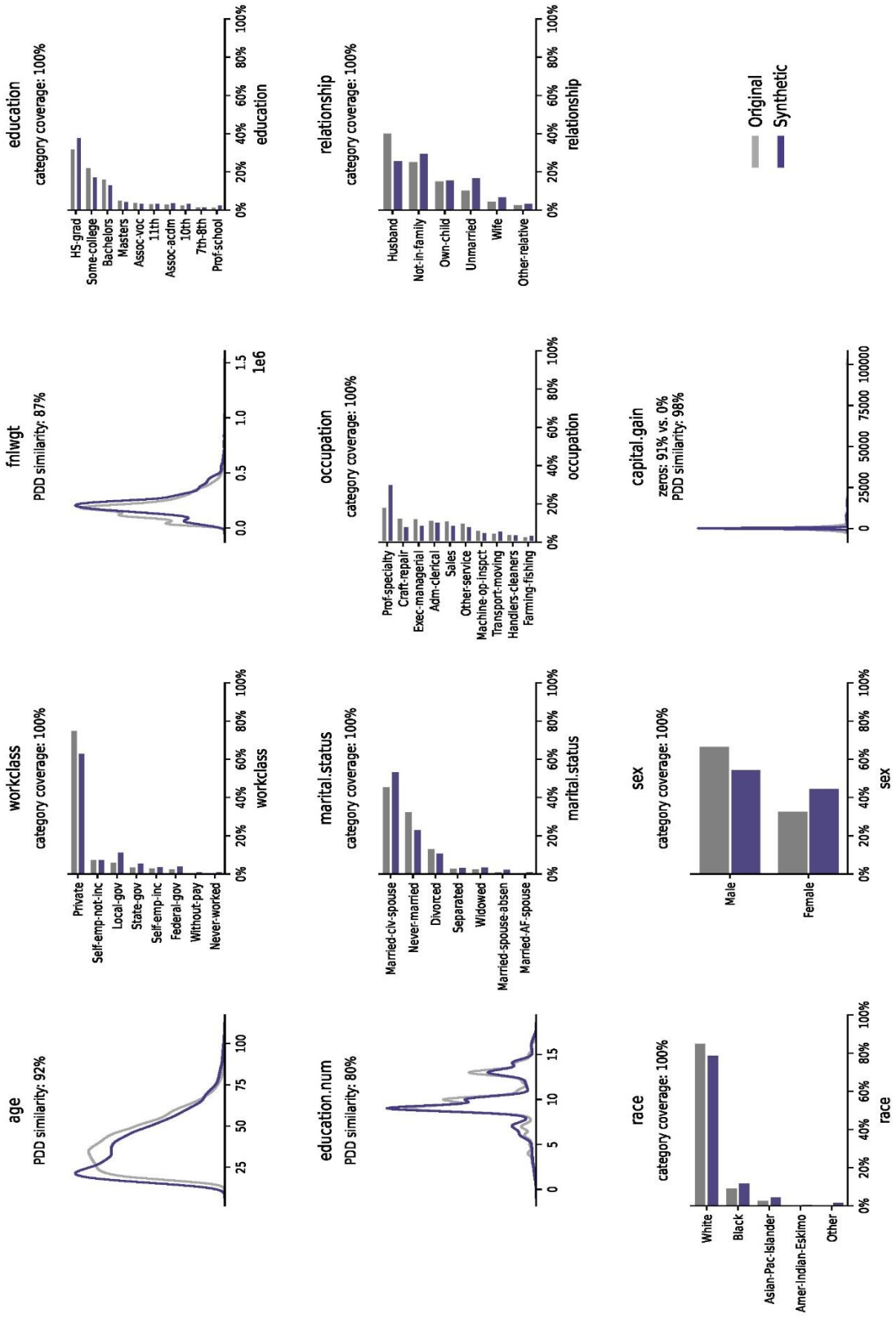

capital.loss
zeros: 95% vs. 1%
PDD similarity: 27%
0
1000
2000
3000
4000
hours.per.week
PDD similarity: 58%
0
20
40
60
80
100
native.country
category coverage: 100%
0%
100%
native.country
income
category coverage: 100%
0%
100%
income
Original
Synthetic

Figure C.2
Univariate distributions with CTGAN



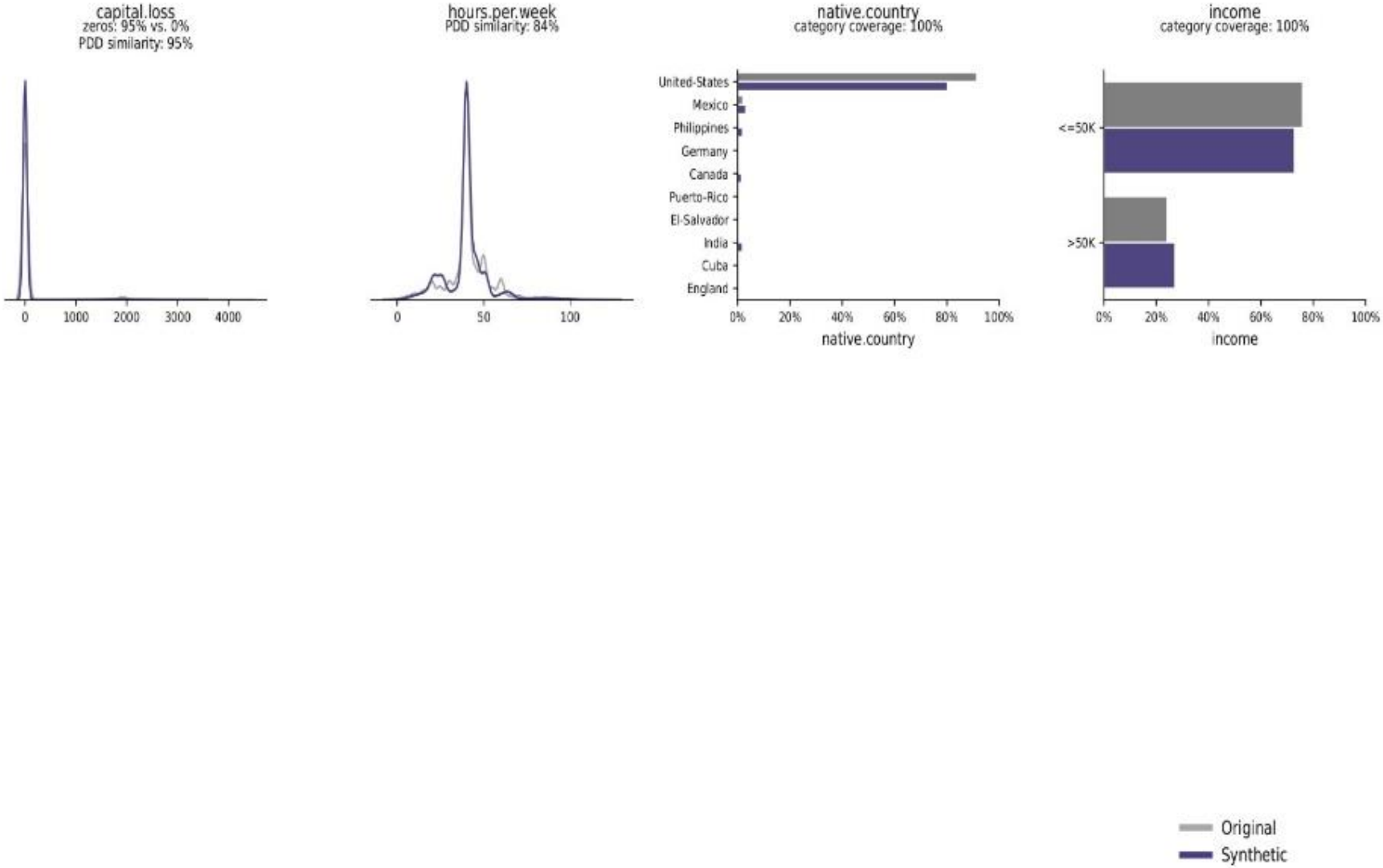
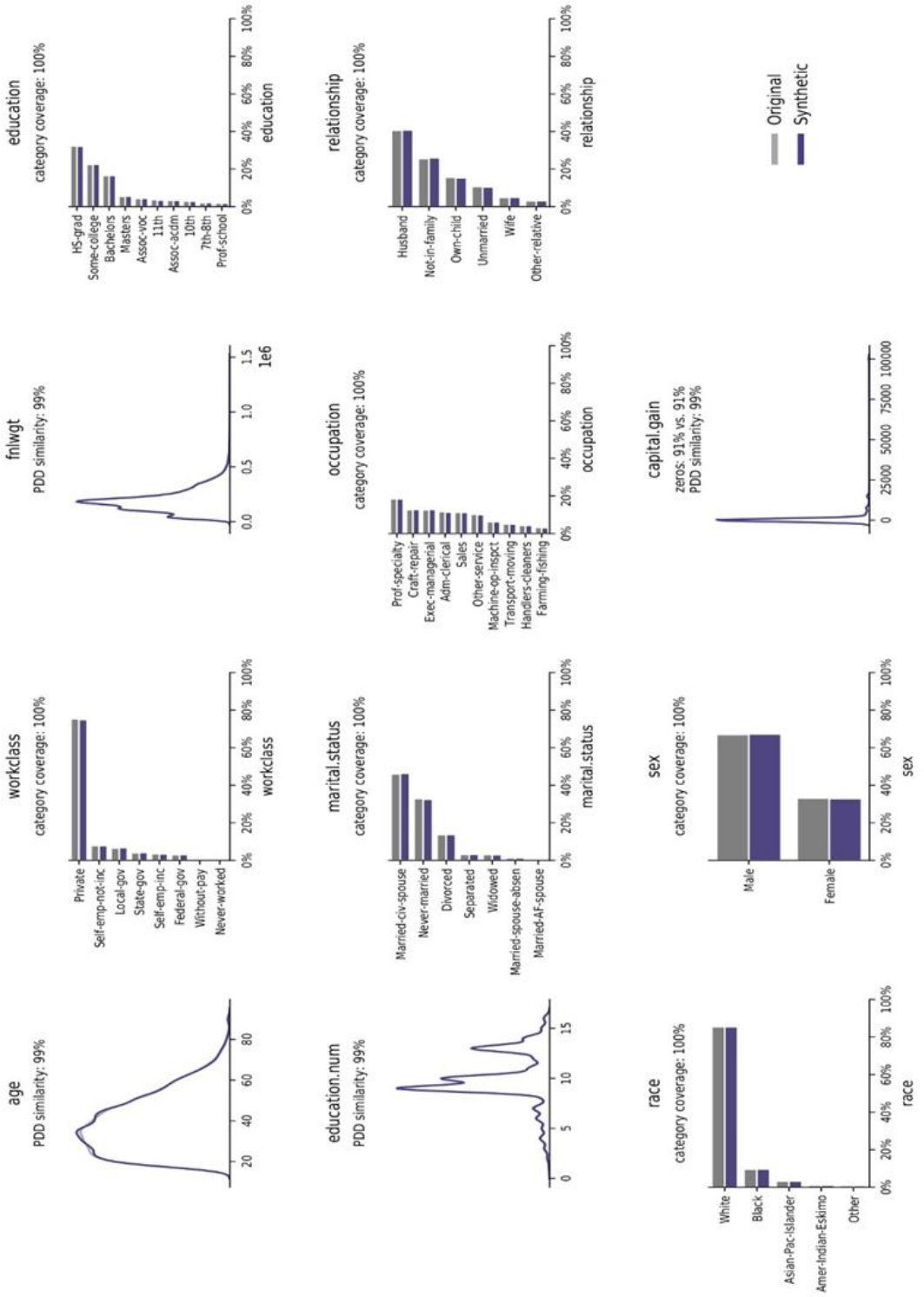
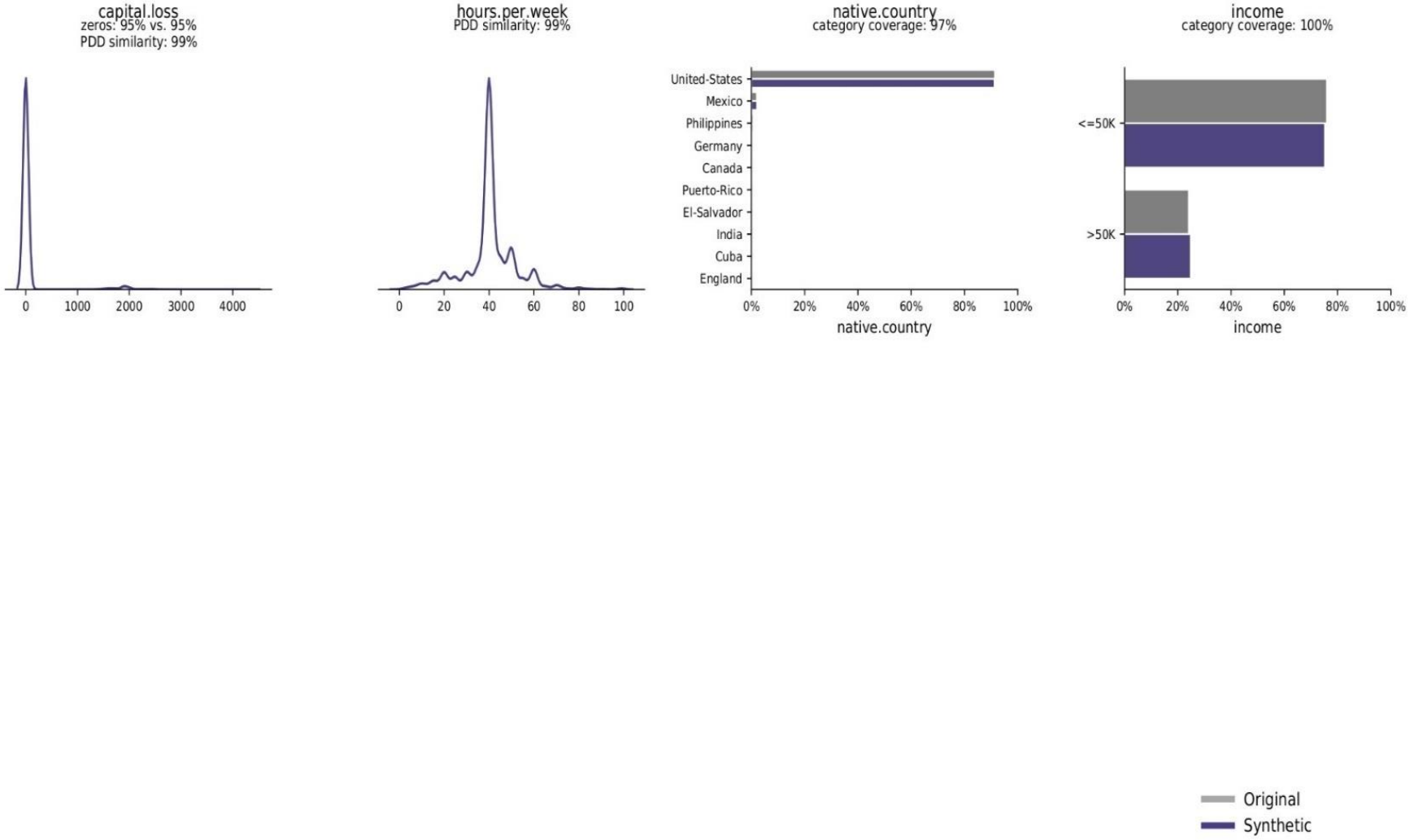

capital.loss
zeros: 95% vs. 0%
PDD similarity: 95%
0
1000
2000
3000
4000
hours.per.week
PDD similarity: 84%
0
50
100
native.country
category coverage: 100%
United-States
Mexico
Philippines
Germany
Canada
Puerto-Rico
El-Salvador
India
Cuba
England
0%
20%
40%
60%
80%
100%
native.country
income
category coverage: 100%
<=50K
>50K
0%
20%
40%
60%
80%
100%
income
Original
Synthetic

Figure C.3
Univariate distributions with SYNTHPOP




capital.loss
zeros: 95% vs. 95%
PDD similarity: 99%
0
1000
2000
3000
4000
hours.per.week
PDD similarity: 99%
0
20
40
60
80
100
native.country
category coverage: 97%
United-States
Mexico
Philippines
Germany
Canada
Puerto-Rico
El-Salvador
India
Cuba
England
0%
20%
40%
60%
80%
100%
native.country
income
category coverage: 100%
<=50K
>50K
0%
20%
40%
60%
80%
100%
income
Original
Synthetic