HAW HAMBURG

BACHELOR THESIS
Jakob Schleiermacher

Evaluierung des PrivBayes-Algorithmus: Vergleich zwischen Original- und synthetischen Daten anhand von Qualitäts- und Datenschutzmetriken

FAKULTÄT TECHNIK UND INFORMATIK
Department Informatik

Faculty of Engineering and Computer Science
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Jakob Schleiermacher

Evaluierung des PrivBayes-Algorithmus: Vergleich zwischen Original- und synthetischen Daten anhand von Qualitäts- und Datenschutzmetriken

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung
im Studiengang *Bachelor of Science Wirtschaftsinformatik*
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr Ulrike Steffens
Zweitgutachter: Prof. Dr. Olaf Zukunft

Eingereicht am: 25.09.2025

Jakob Schleiermacher

Thema der Arbeit

Evaluierung des PrivBayes-Algorithmus: Vergleich zwischen Original- und synthetischen Daten anhand von Qualitäts- und Datenschutzmetriken

Stichworte

Synthetische Daten, Differenzielle Privatsphäre, Datenqualitätsbewertung, Bayessche Netzwerke

Kurzzusammenfassung

In einer zunehmend datengesteuerten Welt bieten synthetische Daten eine praktikable Möglichkeit, hochwertige Testdaten bereitzustellen und gleichzeitig den Schutz personenbezogener Informationen zu gewährleisten. Diese Bachelorarbeit untersucht den PrivBayes-Algorithmus, der Bayessche Netzwerke mit dem Konzept der differenziellen Privatsphäre kombiniert, um synthetische Datensätze zu generieren, die sowohl statistisch aussagekräftig als auch datenschutzkonform sind. Ziel ist es, die Qualität dieser synthetischen Daten anhand zentraler Metriken systematisch zu bewerten.

Dazu werden mehrere öffentlich verfügbare Datensätze verwendet, auf deren Grundlage insgesamt 80 synthetische Varianten unter variierenden Parametereinstellungen entstehen. Die Bewertung erfolgt mithilfe etablierter Metriken wie Weighted Statistics, Column Pair Trends, Value Coverage und Disclosure Protection. Die Ergebnisse zeigen, dass PrivBayes unter bestimmten Bedingungen in der Lage ist, zentrale statistische Strukturen und Variabilität der Originaldaten weitgehend zu erhalten und gleichzeitig ein hohes Maß an Datenschutz zu gewährleisten.

Die Arbeit leistet darüber hinaus einen Beitrag zum Forschungsprojekt INSIGHT (INtelligent Synthesis and Generation of High-quality Test Data), das sich mit der Generierung und Nutzung synthetischer Testdaten befasst. Durch die systematische Evaluation von PrivBayes entsteht ein besseres Verständnis darüber, wie sich Datenschutz und Datenqualität in Einklang bringen lassen. Damit unterstützt die Arbeit das langfristige Ziel der Forschungsgruppe, Werkzeuge und Verfahren zur automatisierten Erzeugung synthetischer Testdaten zu entwickeln, die für realitätsnahe, datenschutzfreundliche Testszenarien geeignet sind.

Jakob Schleiermacher

Title of Thesis

Evaluating PrivBayes: Comparing Original and Synthetic Data with Respect to Quality and Privacy

Keywords

Synthetic Data, Differential Privacy, Data Utility Evaluation, Bayesian Networks

Abstract

In an increasingly data-driven world, synthetic data offers a practical approach to providing high-quality test data while ensuring the protection of personal information. This bachelor thesis examines the PrivBayes algorithm, which combines Bayesian networks with the concept of differential privacy to generate synthetic datasets that are both statistically meaningful and privacy-preserving. The aim is to systematically evaluate the quality of these synthetic datasets using key metrics.

Several publicly available datasets were used as a basis, from which a total of 80 synthetic variants were generated under varying parameter configurations. The evaluation was conducted using established metrics such as Weighted Statistics, Column Pair Trends, Value Coverage, and Disclosure Protection. The results demonstrate that, under specific conditions, PrivBayes can largely preserve the central statistical structures and variability of the original data while maintaining a high level of data protection.

Moreover, this thesis contributes to the INSIGHT (INtelligent Synthesis and Generation of High-quality Test Data) research project, which focuses on the generation and use of synthetic test data. By systematically evaluating PrivBayes, this work helps to deepen the understanding of how privacy and data quality can be effectively balanced. In doing so, it supports the long-term goal of the research group to develop tools and methods for the automated generation of synthetic test data suitable for realistic and privacy-compliant testing scenarios.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Daten spielen heute eine zentrale Rolle in nahezu allen Bereichen von Wirtschaft und Gesellschaft. Sie bilden die Grundlage für datengetriebene Entscheidungen, unterstützen die Entwicklung digitaler Produkte und fördern Innovationen, insbesondere im Bereich der künstlichen Intelligenz [13]. Auch in der Forschung steigt der Bedarf an hochwertigen Daten kontinuierlich, etwa in der Medizin [15, 17] oder in sicherheitskritischen Infrastrukturen [11].

Trotz technologischem Fortschritt und wachsender Datenmengen besteht in vielen Anwendungsfeldern ein erheblicher Mangel an vertrauenswürdigen, rechtlich nutzbaren Datensätzen. Besonders in regulierten Bereichen, wie dem Gesundheits- oder Finanzwesen, erschweren strenge Datenschutzvorgaben und ethische Anforderungen den Zugang zu personenbezogenen Informationen [1, 13, 11]. Kokosi und Harron [15] zeigen am Beispiel klinischer Daten, dass der Zugang häufig mit hohen Kosten und aufwendigen Genehmigungsverfahren verbunden ist. Auch Karst et al. [13] betonen, dass insbesondere kleinere Finanzinstitute durch fehlende Datenverfügbarkeit an der Nutzung moderner KI-Verfahren gehindert werden.

Klassische Schutzmaßnahmen wie Anonymisierung oder Pseudonymisierung gelten zunehmend als unzureichend. Rocher et al. [19] belegen, dass sich selbst stark anonymisierte Datensätze mithilfe von Zusatzwissen erfolgreich rekonstruieren lassen. James et al. [11] zeigen darüber hinaus, dass Anonymisierung häufig mit erheblichem Informationsverlust verbunden ist, wodurch der Nutzen für datenbasierte Anwendungen stark eingeschränkt werden kann. Beduschi [1] weist zudem darauf hin, dass die traditionellen Kategorien „personenbezogen“ und „anonymisiert“ angesichts moderner technischer Entwicklungen zunehmend an Trennschärfe verlieren.

In diesem Kontext gewinnt die Datensynthese zunehmend an Bedeutung. Sie ermöglichet es, künstlich erzeugte Datensätze bereitzustellen, die keine direkten Rückschlüsse auf individuelle Personen zulassen und dennoch zentrale statistische Eigenschaften der Originaldaten bewahren [13, 15]. Qian et al. [17] zeigen, dass synthetische Daten nicht nur für explorative Analysen geeignet sind, sondern auch für Feature-Selektion, Hyperparameter-Optimierung und Modellentwicklung eingesetzt werden können, selbst dann, wenn der Zugriff auf die Originaldaten aus regulatorischen Gründen ausgeschlossen ist. James et al. [11] ergänzen, dass synthetische Daten eine Vielzahl praktischer Anwendungsfelder abdecken, vom internen Softwaretest bis hin zur datenschutzkonformen Weitergabe an externe Partner.

Gerade im Kontext der Testdatengenerierung spielen synthetische Daten eine besondere Rolle. In vielen Softwareprojekten, insbesondere in sicherheitskritischen oder stark regulierten Bereichen, sind realistische Tests unerlässlich. Der Zugriff auf echte Daten ist jedoch oftmals

ausgeschlossen. Ohne qualitativ hochwertige Testdaten lassen sich Systemverhalten oder Fehlertoleranz nicht zuverlässig validieren [11, 17]. Synthetische Daten bieten hier eine praktikable Alternative, um realitätsnahe Testszenarien zu ermöglichen und gleichzeitig datenschutzrechtliche Anforderungen zu erfüllen.

An dieser Stelle setzt das Projekt *INSIGHT* (INtelligent Synthesis and Generation of High-quality Test Data) an, das sich mit der Generierung und Nutzung synthetischer Testdaten für softwaregestützte Anwendungen beschäftigt. Ziel ist es, realitätsnahe, datenschutzkonforme Testdaten bereitzustellen, die eine verlässliche Validierung datenbasierter Systeme ermöglichen. Die hier vorgestellte Untersuchung positioniert sich innerhalb des Teilbereichs „Replikation“ und adressiert die damit verbundenen Herausforderungen bei der Erzeugung, Bewertung und Optimierung synthetischer Daten, die möglichst die Eigenschaften realer Datensätze bewahren sollen.

Ein vielversprechender technischer Ansatz für die Erzeugung synthetischer Daten im Kontext differenzieller Privatsphäre ist das *PrivBayes*-Modell. Dieses Verfahren kombiniert bayessche Netzwerke mit Prinzipien der differenziellen Privatsphäre, um ausgehend von diskretisierten Originaldaten synthetische Datensätze zu erzeugen, die in ihren statistischen Eigenschaften möglichst realitätsnah bleiben und gleichzeitig keine identifizierbaren Einzelinformationen preisgeben [30].

In der bisherigen Forschung wurde PrivBayes bereits in zahlreichen Arbeiten eingesetzt und evaluiert. Dabei kommen typischerweise etablierte Metriken wie der Kolmogorov-Smirnov-Test für kontinuierliche Merkmale oder der Chi-Quadrat-Test für kategoriale Merkmale zum Einsatz, um die Übereinstimmung statistischer Strukturen zwischen realen und synthetischen Daten zu bewerten [4, 16]. Diese Verfahren erlauben eine grundlegende Einschätzung der Datenqualität, konzentrieren sich jedoch häufig auf univariate oder marginale Verteilungen und lassen komplexere Zusammenhänge unbeachtet[30, 4, 16].

Aktuelle Studien deuten darauf hin, dass diese klassischen Bewertungsmethoden in komplexeren Anwendungskontexten oft nicht ausreichen, um die tatsächliche Eignung synthetischer Daten fundiert zu beurteilen. Sarmin et al. [21], Du und Li [6] sowie Sella et al. [22] fordern eine breitere Perspektive auf die Evaluation synthetischer Daten. Neben der statistischen Ähnlichkeit sollten auch strukturelle, semantische und generative Eigenschaften berücksichtigt werden, um die Nutzbarkeit in praxisnahen Szenarien realistisch einschätzen zu können.

Aufbauend auf diesen Überlegungen verfolgt die vorliegende Arbeit das Ziel, die Qualität synthetischer Datensätze, die mit PrivBayes auf Basis öffentlich verfügbarer Originaldatensätze erzeugt wurden, systematisch zu bewerten. Hierzu werden synthetische und Originaldaten systematisch anhand ergänzender Metriken gegenübergestellt, um ein differenziertes Bild der Datenqualität zu gewinnen und die grundsätzliche Eignung synthetischer Daten für Test- und Validierungsszenarien zu beurteilen. Eine direkte Validierung im Rahmen produktiver Softwaretests oder Machine-Learning-Anwendungen erfolgt in dieser Arbeit nicht.

Zur Umsetzung dieses Ziels werden Experimente mit mehreren öffentlich verfügbaren Datensätzen durchgeführt. Die Auswahl und Anwendung spezifischer Metriken ermöglichen eine strukturierte Analyse der Auswirkungen verschiedener PrivBayes-Parametrisierungen auf die resultierende Datenqualität.

Diese Arbeit untersucht, inwiefern synthetische Daten, die mit PrivBayes unter Wahrung differenzieller Privatsphäre erzeugt wurden, für realitätsnahe Test- und Validierungsszenarien geeignet sind. Der Fokus liegt auf einer systematischen Metrikanalyse entlang der Dimensionen statistische Ähnlichkeit, Variabilität und Datenschutz. Ergänzend wird die Leistungsfähigkeit von PrivBayes in Relation zu CTGAN und TVAE eingeordnet.

Auf dieser Grundlage ergeben sich die folgenden zentralen Forschungsfragen:

1. Wie gut erhält PrivBayes die statistischen Eigenschaften der Originaldaten?
2. Wie gut reproduziert PrivBayes die Variabilität der Originaldaten?
3. Inwieweit gewährleistet PrivBayes den Schutz sensibler Informationen?
4. Gibt es eine Konfiguration[1], die über alle Datensätze hinweg gute Ergebnisse erzielt?
5. Wie unterscheiden sich die Ergebnisse von PrivBayes, CTGAN und TVAE?

Die Arbeit ist wie folgt aufgebaut: Kapitel 2 stellt die theoretischen Grundlagen vor, darunter bayessche Netzwerke, differenzielle Privatsphäre und das Konzept synthetischer Daten. Kapitel 3 erläutert das experimentelle Vorgehen sowie die eingesetzten Bewertungsmethoden. Im Anschluss werden in Kapitel 4 die Resultate präsentiert und im Kapitel 5 kritisch diskutiert. Kapitel 6 fasst die zentralen Erkenntnisse zusammen und gibt Ideen für weiterführende Arbeiten.

[1] Details zur Konfiguration des verwendeten Algorithmus werden im Methodik-Kapitel erläutert.

2 Grundlagen

In diesem Kapitel werden die zentralen theoretischen und methodischen Grundlagen dieser Arbeit erläutert. Ziel ist es, ein grundlegendes Verständnis für die verwendeten Konzepte und Verfahren zu vermitteln und den fachlichen Kontext nachvollziehbar darzustellen. Die folgenden Abschnitte bilden die Basis für die späteren Analysen zur Bewertung synthetischer Daten.

Zunächst erfolgt eine Einführung in das Konzept synthetischer Daten, das sich als vielversprechender Ansatz zur datenschutzfreundlichen Nutzung sensibler Informationen etabliert hat. Darauf aufbauend werden grundlegende Verfahren generativer Modelle vorgestellt, wobei die Modelle CTGAN und TVAE exemplarisch behandelt werden. Diese dienen im weiteren Verlauf als Vergleich zu dem in dieser Arbeit untersuchten Verfahren.

Anschließend werden die Konzepte erläutert, die dem Modell PrivBayes zugrunde liegen. Dazu zählen insbesondere Bayessche Netzwerke als Grundlage für die Modellierung von Wahrscheinlichkeitsverteilungen sowie das Prinzip der differenziellen Privatsphäre zur Wahrung individueller Datenschutzanforderungen. Abschließend wird PrivBayes vorgestellt, das beide Ansätze kombiniert und eine Methode zur Generierung synthetischer Datensätze bietet, die sowohl statistisch aussagekräftig als auch datenschutzkonform sind.

In der vorliegenden Arbeit wird PrivBayes verwendet, um synthetische Datensätze auf Basis öffentlich verfügbarer Demo-Daten zu erzeugen. Dabei handelt es sich nicht um produktive Testdaten aus realen Systemen, sondern um repräsentative Datensätze, die eine experimentelle Bewertung unter kontrollierten Bedingungen ermöglichen. Ziel ist es, die Qualität der erzeugten Daten zu analysieren und daraus Rückschlüsse auf deren potenzielle Eignung für realitätsnahe Testszenarien zu ziehen.

2.1 Synthetische Daten: Motivation, Einsatz und Bedeutung in der Softwareentwicklung

In der Softwareentwicklung besteht zunehmend die Herausforderung, realistische Testdaten für große, sensible Datensätze zu generieren und dabei gleichzeitig Datenschutzvorgaben wie die DSGVO oder HIPAA einzuhalten. Synthetische Daten gelten in diesem Kontext als praktikabler Kompromiss, da sie die Verteilungseigenschaften realer Datensätze nachbilden, ohne personenbezogene Informationen offenzulegen [10].

Dabei wird auf Grundlage eines bestehenden Datensatzes ein Modell trainiert, das die relevanten Zusammenhänge zwischen den Attributen abbildet. Anschließend erzeugt dieses Modell neue Daten, die in Struktur und Verteilung ähnlich sind, aber keine realen Einträge enthalten [10]. Neben der formalen Trennung ist entscheidend, wie gut die zugrunde liegenden Abhängigkeiten reproduziert werden. Die Balance zwischen Realitätsnähe und Datenschutz erfordert dabei eine kontextabhängige Abwägung.

Insbesondere in der Qualitätssicherung werden synthetische Daten relevant. Die manuelle Erstellung von Testdaten oder die Anonymisierung der Produktivdaten sind oft mit erheblichem Aufwand verbunden [26, 24]. Die manuelle Modellierung komplexer Zusammenhänge ist ressourcenintensiv, insbesondere bei vielen Variablen [26]. Anonymisierung führt zudem nicht immer zu sicherem Datenschutz [26, 24]. Synthetische Testdaten bieten hier Vorteile: Sie lassen sich automatisiert generieren, besser an gesetzliche Vorgaben anpassen und gezielt hinsichtlich der Testabdeckung erzeugen.

Weitere Vorteile liegen in der gezielten Erzeugung seltener, sicherheitskritischer Szenarien und der Wiederverwendbarkeit im Rahmen iterativer Entwicklungsprozesse [8, 10]. Gerade in agilen Umgebungen ermöglichen synthetische Daten eine flexible Anpassung an neue Anforderungen und unterstützen die Entwicklung skalierbarer Teststrategien.

Gleichzeitig hängt die Qualität synthetischer Daten stark vom gewählten Verfahren, den Modellannahmen und den Trainingsdaten ab. Fehlerhafte Modellierung kann zu unplausiblen Werten oder sogar zu Datenschutzproblemen führen [10]. Auch in der Praxis zeigen sich Probleme, wenn ohne geeignete Strukturierungs- und Validierungsregeln synthetische Daten entstehen, die für Tests ungeeignet sind [26]. Die Auswahl eines geeigneten Generierungsverfahrens ist daher entscheidend.

In dieser Arbeit wird untersucht, wie sich unterschiedliche Konfigurationen des PrivBayes-Modells auf datenbasierte Metriken auswirken. Ziel ist es, zentrale Parameter hinsichtlich ihrer Wirkung auf die Qualität synthetisch erzeugter Demo-Datensätze zu evaluieren und deren potenzielle Eignung für realitätsnahe Testdaten abzuleiten.

2.2 Modellklassen generativer Verfahren: Statistikbasierte und neuronale Ansätze

Generative Modelle zur Datensynthese lassen sich grob in zwei Klassen unterteilen: statistikbasierte Verfahren, die auf expliziten Annahmen über Verteilungen und Abhängigkeitsstrukturen beruhen, und datengetriebene Verfahren auf Basis neuronaler Netze, die komplexe Zusammenhänge direkt aus Trainingsdaten lernen.

Auch statistische Modelle wie PrivBayes lernen aus Trainingsdaten, allerdings auf Grundlage eines expliziten Modells: Sie analysieren gezielt Verteilungen und Abhängigkeiten zwischen Attributen und integrieren Datenschutzmechanismen wie Differenzielle Privatsphäre in diesen Prozess. Ihr Vorteil liegt in der transparenten Modellstruktur und der gezielten Kontrolle

der zugrunde liegenden Annahmen. Allerdings können sie tiefere oder implizite Strukturen in den Daten nur eingeschränkt erfassen. Neuronale Modelle wie CTGAN und TVAE sind flexibler, können hochdimensionale, gemischt-typige Daten abbilden, sind jedoch auf große Datenmengen angewiesen und schwerer interpretierbar [28].

In dieser Arbeit steht PrivBayes im Mittelpunkt. Um dessen Ergebnisse kontextualisieren zu können, werden CTGAN und TVAE als neuronale Vergleichsmodelle herangezogen. CTGAN basiert auf Generative Adversarial Networks und erweitert diese um eine konditionierte Trainingsstrategie und eine mode-spezifische Normalisierung zur besseren Abbildung multimodaler Verteilungen. TVAE hingegen ist ein variationaler Autoencoder, der so angepasst wurde, dass gemischte Datentypen über geeignete Verlustfunktionen verarbeitet werden können [28].

2.3 Bayessche Netzwerke als Grundlage probabilistischer Modellierung

Bayessche Netzwerke (BNs) sind gerichtete azyklische Graphen, die gemeinsame Wahrscheinlichkeitsverteilungen kompakt darstellen. Jeder Knoten entspricht einer Zufallsvariablen, die Kanten kodieren bedingte Abhängigkeiten [12].

Die zentrale Eigenschaft ist die Zerlegung der gemeinsamen Verteilung $P(X_1, X_2, ..., X_n)$ in ein Produkt bedingter Wahrscheinlichkeiten:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i))$$

Diese Zerlegung basiert auf dem Konzept der bedingten Unabhängigkeit und erlaubt effiziente Berechnung und Speicherung auch für große Datensätze [12].

Bayessche Netzwerke ermöglichen eine transparente und strukturierte Repräsentation von Domänenwissen. Sie lassen sich gut interpretieren, sind erweiterbar und können zur Generierung neuer Daten verwendet werden, indem aus der gelernten Verteilung gesampelt wird. Dieses Prinzip nutzt PrivBayes zur Erzeugung synthetischer Daten [30].

2.4 Differenzielle Privatsphäre als Datenschutzprinzip

Differenzielle Privatsphäre (DP) ist ein formales Datenschutzkonzept, das sicherstellt, dass die Einbeziehung oder Entfernung eines einzelnen Eintrags den Ausgang eines Analyseverfahrens nur unwesentlich beeinflusst. Die Privatsphäre wird über den Parameter ε quantifiziert: Kleine Werte bedeuten hohen Datenschutz, große Werte höhere Genauigkeit. DP ist anwendbar, unabhängig von Datenstruktur oder Angreiferwissen. In der synthetischen Datengenerierung wird DP oft durch das Hinzufügen von Rauschen zu Aggregationen oder Modellparametern realisiert. Dadurch entstehen Daten, die strukturell ähnlich, aber nicht auf einzelne Personen

zurückführbar sind. Die Wahl des ε-Werts stellt dabei stets einen Kompromiss zwischen Datenschutz und Nutzbarkeit dar. PrivBayes zeigt exemplarisch, wie DP in ein Modell integriert werden kann: Die Privatsphäre wird bereits beim Lernen der Modellstruktur abgesichert, nicht erst bei der Ausgabe der Daten [7, 30].

2.5 PrivBayes: Synthetische Datengenerierung unter Wahrung der Privatsphäre

PrivBayes kombiniert bayessche Netzwerke mit dem Konzept der Differenziellen Privatsphäre, um synthetische Datensätze zu erzeugen, die sowohl die statistischen Zusammenhänge der Originaldaten bewahren als auch formale Datenschutzanforderungen erfüllen [30].

Das Verfahren besteht aus mehreren Schritten:

Zunächst wird die Struktur des bayesschen Netzwerks gelernt. Dieser Schritt wird als Strukturlernen bezeichnet und legt fest, in welcher Reihenfolge die Attribute zueinander in Beziehung stehen. In jedem Schritt wird das Attribut mit der höchsten Mutual Information zu den bereits gewählten Attributen ausgewählt. Mutual Information misst, wie stark zwei Variablen statistisch voneinander abhängen. Um die Privatsphäre zu wahren, wird diese Auswahl nicht deterministisch getroffen. PrivBayes verwendet hierfür den Exponential Mechanismus, der informativen Attributen zwar eine höhere Auswahlwahrscheinlichkeit zuweist, aber auch bewusst weniger optimale Entscheidungen zulässt, um Rückschlüsse auf einzelne Datenpunkte zu erschweren.

Im nächsten Schritt folgt die Parameterschätzung. Hierbei werden die bedingten Wahrscheinlichkeiten zwischen den Attributen geschätzt, also etwa: Wie wahrscheinlich ist ein bestimmter Wert eines Attributs, wenn ein anderer bereits bekannt ist? Um die Privatsphäre dabei zu wahren, wird der Laplace Mechanismus eingesetzt. Dieser fügt gezielt Rauschen zu den gezählten Häufigkeiten hinzu, sodass keine genauen Rückschlüsse auf einzelne Einträge im ursprünglichen Datensatz möglich sind.

Auf Basis der gelernten Struktur und der geschätzten Wahrscheinlichkeiten wird schließlich ein synthetischer Datensatz erzeugt. Dieser Schritt erfolgt über Bayesian Network Sampling, bei dem neue Datenzeilen gezogen werden, die den modellierten Abhängigkeiten folgen, aber keine echten Daten enthalten.

Durch die modulare Architektur lässt sich der Datenschutz gezielt steuern. Das gesamte Privacy-Budget ε wird auf die beiden Hauptkomponenten Strukturlernen und Parameterschätzung verteilt. PrivBayes eignet sich aufgrund seiner transparenten Modellierung besonders für Anwendungen mit hohen Anforderungen an Datenschutz und Nachvollziehbarkeit [30].

Beispielhafte Veranschaulichung von PrivBayes Zur Veranschaulichung des Verfahrens wird ein kleiner fiktiver Datensatz betrachtet, der Personen anhand dreier Merkmale beschreibt: Alter (jung, alt), Einkommen (niedrig, hoch) und Versicherung (ja, nein). PrivBayes durchläuft mehrere Schritte, um daraus ein bayessches Netzwerk zu lernen, das die statistischen Abhängigkeiten zwischen den Attributen beschreibt.

1. **Strukturlernen:** PrivBayes analysiert, welche Attribute statistisch stark miteinander zusammenhängen. Dazu wird die sogenannte Mutual Information (MI) zwischen den Attributen berechnet. Angenommen, die höchste MI besteht zwischen Alter und Einkommen. Der Exponential Mechanismus sorgt nun dafür, dass dieses Attributpaar mit hoher Wahrscheinlichkeit als Kante in das bayessche Netzwerk aufgenommen wird, allerdings nicht deterministisch, sondern zufallsbasiert zum Schutz der Privatsphäre.

2. **Parameterschätzung:** Auf Basis der gewählten Struktur werden die bedingten Wahrscheinlichkeiten geschätzt, etwa:

$$P(\text{Versicherung} = \text{ja} \mid \text{Alter} = \text{jung}, \text{Einkommen} = \text{hoch}).$$

   Um die Privatsphäre zu wahren, wird bei der Schätzung der Wahrscheinlichkeiten der Laplace Mechanismus eingesetzt, der gezielt Rauschen hinzufügt.

3. **Datengenerierung:** Mithilfe des erlernten Netzwerks und der geschätzten Wahrscheinlichkeiten werden neue Datenzeilen erzeugt, die statistisch konsistent mit den ursprünglichen Verteilungen sind, aber keine echten Personen mehr repräsentieren.

Dieses Beispiel zeigt, wie PrivBayes gleichzeitig statistische Muster bewahren und datenschutzkonform synthetische Daten generieren kann.

3 Methodik

Dieses Kapitel beschreibt das methodische Vorgehen zur Erzeugung und Bewertung synthetischer Daten. Ziel ist es, die eingesetzten Datensätze, die notwendigen Vorverarbeitungsschritte, die Konfiguration des Generierungsverfahrens sowie die verwendeten Werkzeuge und Vergleichsmetriken nachvollziehbar darzustellen.

Zunächst werden die verwendeten Datensätze und deren wesentliche Eigenschaften erläutert. Ein besonderes Augenmerk liegt dabei auf den vorbereitenden Schritten (Preprocessing), die im Rahmen von PrivBayes automatisch durchgeführt werden. Auch wenn dieser Prozess technisch im Hintergrund erfolgt, ist es wichtig zu verstehen, welche Transformationen dabei vorgenommen werden, da diese eine zentrale Voraussetzung dafür bilden, dass die erzeugten synthetischen Daten überhaupt sinnvoll nutzbar und qualitativ bewertbar sind.

Im Anschluss wird die Konfiguration von PrivBayes vorgestellt. Neben dem Datenschutzwert ε werden dabei insbesondere die gewählte Anzahl an Bins sowie die Begrenzung der maximal zulässigen Anzahl an abhängigen Elternknoten im Bayesschen Netzwerk beschrieben. Außerdem wird die Python-Bibliothek Synthcity kurz erläutert, deren bestehende Implementierung von PrivBayes in dieser Arbeit verwendet wurde.

Abschließend wird das Vorgehen zur Bewertung der synthetischen Datenqualität beschrieben. Dazu werden die verwendeten Vergleichsmetriken vorgestellt, anhand derer die erzeugten Daten mit den Originaldaten verglichen werden.

3.1 Datenbasis

Für die Evaluation der synthetischen Datengenerierung wurden verschiedene Datensätze verwendet, die sich hinsichtlich Umfang, Struktur und inhaltlichem Kontext unterscheiden. Ziel dabei ist es, die Funktionsweise und die Qualität von PrivBayes unter möglichst unterschiedlichen Rahmenbedingungen zu überprüfen. Die Datensätze stammen aus öffentlich verfügbaren Quellen und decken verschiedene Anwendungsbereiche ab, darunter sozioökonomische, medizinische, versicherungstechnische sowie umwelt- und nachrichtenbezogene Daten.

Eine Übersicht der verwendeten Datensätze ist in Tabelle 3.1 dargestellt. Sie enthalten sowohl kontinuierliche (Kon) als auch kategoriale (Kat) Merkmale, sind multivariat aufgebaut und weisen keine fehlenden Werte auf.

Aufgrund begrenzter Rechenkapazität wurden die Datensätze *Covtype* und *News* vor der Analyse stichprobenartig reduziert. Dabei steht eine einfache Klammer `( )` in der Tabelle für eine Reduzierung auf 1.000 Zeilen und eine doppelte Klammer `(( ))` für eine Reduzierung auf 500 Zeilen. Die Ziehung der Stichproben erfolgte zufällig, wobei durch Setzen des Parameters `random_state = 42` eine Reproduzierbarkeit gewährleistet wurde.

Tabelle 3.1: Übersicht der verwendeten Datensätze

Name	**Domäne**	**Zeilen**	**Attribute**	**Kon.**	**Kat.**	**Quelle**
Adult	Soziodemografisch	32.561	15	6	9	[14]
Child	Medizinisch	20.000	20	0	20	[25]
Covtype	Umwelt	(5.000)	55	55	0	[3]
Insurance	Versicherung	20.000	27	0	27	[2]
News	Nachrichtenwirtschaft	((5.000))	59	59	0	[9]

3.1.1 Vorbereitende Transformationen im Rahmen von PrivBayes

PrivBayes ist auf die Verarbeitung kategorialer Daten ausgelegt. Um auch kontinuierliche Merkmale verwenden zu können, werden diese vor der Modellierung durch Diskretisierung in feste Intervalle unterteilt. Die so entstehenden Bins repräsentieren kategoriale Ausprägungen der ursprünglich kontinuierlichen Attribute.

Bereits kategorial vorliegende Merkmale können grundsätzlich direkt verwendet werden, müssen jedoch intern numerisch kodiert werden. Hierzu kommen beispielsweise einfache Ganzzahlzuweisungen oder binäre Kodierungen zum Einsatz, die keine inhaltliche Veränderung der Daten darstellen, sondern lediglich die technische Verarbeitung im Rahmen des Algorithmus ermöglichen. Da PrivBayes keine Mechanismen zur Behandlung fehlender Werte vorsieht, wurde vor der Analyse sichergestellt, dass in den verwendeten Datensätzen keine Werte fehlen. Die genannten Transformationen werden vom Algorithmus im Rahmen des Modellaufbaus automatisiert durchgeführt.

3.2 Einsatz von Synthcity und Konfiguration von PrivBayes

Zur Generierung der synthetischen Datensätze wurde das Python-Framework Synthcity verwendet, dass eine einheitliche Schnittstelle zur Anwendung verschiedener generativer Verfahren für strukturierte Tabellendaten bereitstellt [18].

In dieser Arbeit kam dabei ausschließlich das Modell PrivBayes innerhalb von Synthcity zum Einsatz. Vergleichsdaten für alternative Generierungsverfahren wurden von der begleitenden Forschungsgruppe im Rahmen des Projekts INSIGHT bereitgestellt.

Zur Erzeugung der synthetischen Datensätze mit PrivBayes wurden unterschiedliche Konfigurationen getestet, um die Auswirkungen zentraler Modellparameter auf die Datenqualität

und den Datenschutz systematisch zu untersuchen. Dabei wurden folgende Parameter berücksichtigt:

Die Anzahl der Bins, mit denen kontinuierliche Attribute vor der Modellierung in diskrete Intervalle unterteilt werden, ist ein technisch notwendiger Bestandteil der Datentransformation in PrivBayes. Konkrete Empfehlungen oder systematische Analysen zum Einfluss dieses Parameters finden sich in der bisherigen Literatur zu PrivBayes nicht. Daher wurden im Rahmen dieser Arbeit die technisch gängigen Konfigurationen mit 10 und 20 Bins verwendet, um eine praxisnahe Diskretisierung sicherzustellen.

Ein zentraler Parameter ist die maximale Anzahl erlaubter Elternknoten k pro Attribut innerhalb des Bayesschen Netzwerks. Dieser Wert steuert die Komplexität der modellierten Abhängigkeitsstrukturen. Höhere Werte erlauben es dem Modell, komplexere Zusammenhänge zwischen den Attributen abzubilden. Gleichzeitig steigt mit zunehmendem k jedoch auch die Modellkomplexität, was je nach Datensatz die Interpretierbarkeit erschweren und das Risiko einer Überanpassung erhöhen kann. In früheren Arbeiten wurde k daher üblicherweise niedrig gewählt, um ein ausgewogenes Verhältnis zwischen Modellflexibilität und Robustheit sicherzustellen. Aufbauend auf diesen Erkenntnissen wurden in dieser Arbeit die Werte $k = 2$ und $k = 3$ getestet.

Der Datenschutzparameter ε bestimmt das Maß an erlaubtem Informationsverlust im Sinne der Differenziellen Privatsphäre. Die gewählten Werte 0,1, 0,5, 1,0 und 10,0 orientieren sich an gängigen Konfigurationen aus der Literatur und decken ein breites Spektrum an Datenschutzstufen ab, von starkem Schutz bei niedrigen ε-Werten bis hin zu datenähnlicher Reproduktion bei höheren Werten.

Durch die Kombination der gewählten Konfigurationswerte ergeben sich insgesamt $2 \times 2 \times 4 = 16$ verschiedene synthetische Datensätze pro Originaldatensatz. Über alle fünf verwendeten Datensätze hinweg wurden somit insgesamt 80 synthetische Datensätze generiert. Die konkrete Umsetzung der Datengenerierung erfolgte über die Schnittstellen von Synthcity. Ein exemplarischer Python-Code zur Generierung synthetischer Daten mit PrivBayes ist im Anhang A dieser Arbeit enthalten und veranschaulicht das Vorgehen.

3.3 Bewertungskonzept und eingesetzte Metriken

Die Qualität der in dieser Arbeit erzeugten synthetischen Datensätze wurde anhand etablierter Metriken bewertet, die eine differenzierte Einschätzung der Datenähnlichkeit, Variabilität und Datenschutzqualität ermöglichen. Zur technischen Umsetzung kam der Evaluation Service der HAW Hamburg zum Einsatz, der auf die Open-Source-Bibliothek SDMetrics zurückgreift [5]. SDMetrics bietet eine Vielzahl von Metriken zur quantitativen Bewertung synthetischer Daten, die in aktuellen Forschungsarbeiten und Projekten Verwendung finden. Die Ausgabewerte der Metriken sind dabei so skaliert, dass ein Wert von 0 die schlechteste und ein Wert von 1 die bestmögliche Ausprägung repräsentiert.

Um die Bewertung systematisch entlang der in dieser Arbeit definierten Forschungsfragen auszurichten, wurden die verfügbaren Metriken inhaltlich strukturiert. Auf dieser Grundlage entstand ein dreistufiges Bewertungsschema, das die Metriken den Aspekten statistische Ähnlichkeit, Variabilität und Datenschutz zuordnet. Für die Kategorie statistische Ähnlichkeit kommen etablierte Verfahren zum Einsatz, die sich in der quantitativen Bewertung synthetischer Daten bewährt haben. Die Einordnung der übrigen Metriken in die Kategorien Variabilität und Datenschutz erfolgte im Rahmen dieser Arbeit, um den von Sarmin et al. [21], Du und Li [6] sowie Sella et al. [22] geforderten erweiterten Blickwinkel auf die Qualität synthetischer Daten konzeptionell zu berücksichtigen.

Statistische Struktur

Zur Beurteilung der statistischen Struktur synthetischer Daten wurde untersucht, inwieweit zentrale Merkmale des ursprünglichen Datensatzes erhalten bleiben. Dabei lag der Fokus auf univariaten Verteilungen, Attributbeziehungen sowie der Einhaltung von Wertebereichen. Diese Aspekte wurden anhand der folgenden drei Metriken erfasst:

Weighted Statistics setzt sich aus zwei Teilmetriken zusammen: *KSComplement* für numerische Attribute und *TVComplement* für kategoriale Merkmale. Beide Teilmetriken vergleichen die Verteilung eines Attributs im Original- und im synthetischen Datensatz und bewerten deren Ähnlichkeit. *KSComplement* verwendet dabei den Kolmogorov-Smirnov-Test, um Unterschiede in den kumulativen Verteilungen numerischer Spalten zu identifizieren. *TVComplement* hingegen nutzt die Total Variation Distance, um die Verteilungsunterschiede kategorialer Attribute zu messen. Beide Werte werden so transformiert, dass ein hoher Score auf eine große Übereinstimmung der Verteilungen hinweist. Die **Weighted Statistics**-Metrik kombiniert beide Teilmetriken zu einem gewichteten Gesamtscore, der die univariate Ähnlichkeit über alle Attribute hinweg zusammenfasst [5].

Boundary Adherence prüft, ob die synthetischen Daten die in den Originaldaten beobachteten Wertebereiche einhalten. Dafür wird analysiert, welcher Anteil der Werte in den synthetischen Daten innerhalb der Minimal- und Maximalwerte der Originaldaten liegt. Werte außerhalb dieses Intervalls fließen negativ in die Bewertung ein. Ein hoher Score bedeutet, dass die synthetischen Daten keine Ausreißer über die beobachteten Grenzen hinaus enthalten und damit eine realitätsnahe Werteverteilung aufweisen [5].

Column Pair Trends analysieren, ob die Beziehungen zwischen Attributpaaren im synthetischen Datensatz erhalten geblieben sind. Abhängig vom Datentyp der verglichenen Spalten kommen unterschiedliche Verfahren zum Einsatz. Für numerische Spaltenpaare wird die *CorrelationSimilarity* verwendet, die prüft, wie stark sich Korrelationskoeffizienten zwischen realen und synthetischen Daten ähneln. Für kategoriale Spaltenpaare wird die *ContingencySimilarity* herangezogen, die auf dem Vergleich von Kontingenztabellen basiert. Für gemischte Paare aus numerischen und kategorialen Attributen wird der numerische Wert zunächst diskretisiert, bevor ebenfalls die *ContingencySimilarity* angewendet wird. Der endgültige Wert

der **Column Pair Trends** ergibt sich aus dem Durchschnitt der Einzelbewertungen aller Attributpaare [5].

Insgesamt ermöglichen diese drei Metriken eine fundierte Einschätzung darüber, ob die generierten Daten grundlegende statistische Eigenschaften des Originals bewahren. Aufbauend auf dieser Bewertung folgt im nächsten Abschnitt die Analyse der Variabilität der synthetischen Daten.

Variabilität

Die Kategorie Variabilität bewertet die Vielfalt innerhalb der synthetischen Daten. Ziel ist es, zu untersuchen, ob die generierten Daten sowohl neue Ausprägungen enthalten als auch die ursprüngliche Bandbreite der Originaldaten abdecken. Zur Bewertung werden die Metriken **Value Coverage** und **New Row Synthesis** herangezogen.

Value Coverage setzt sich aus zwei Komponenten zusammen: *Category Coverage* für kategoriale Daten und *Range Coverage* für numerische Daten.

Category Coverage ermittelt den Anteil der Kategorien in den synthetischen Daten, die auch in den Originaldaten vorkommen. Ein niedriger Wert weist darauf hin, dass im synthetischen Datensatz bestimmte Kategorien des Originals nicht repräsentiert wurden. Da PrivBayes ausschließlich mit den im Original vorhandenen Kategorien arbeitet und keine neuen Werte erzeugt, ist ein niedriger Wert in diesem Kontext stets auf fehlende, nicht generierte Kategorien zurückzuführen.

Range Coverage bewertet, ob der numerische Wertebereich der Originaldaten durch die synthetischen Daten vollständig abgedeckt wird. Dabei wird geprüft, inwieweit sich die minimalen und maximalen Werte der synthetischen Daten den entsprechenden Grenzen der Originaldaten annähern und ob innerhalb dieses Bereichs keine Lücken entstehen. Fehlende Zwischenbereiche, etwa wenn bestimmte Werte im mittleren Bereich des Originals nicht synthetisch erzeugt wurden, führen zu einer Abwertung.

Beide Teilmetriken ignorieren die Häufigkeit einzelner Werte und betrachten ausschließlich deren Vorkommen. Ein hoher **Value Coverage**-Wert signalisiert, dass die synthetischen Daten die Vielfalt der Originaldaten in Bezug auf Wertebereiche und Kategorien angemessen widerspiegeln [5].

New Row Synthesis quantifiziert, in welchem Maß die erzeugten synthetischen Daten neue Datenzeilen darstellen, die nicht exakt mit Zeilen aus dem Originaldatensatz übereinstimmen. Eine Übereinstimmung liegt vor, wenn sämtliche Werte einer synthetischen Zeile mit den Werten einer Originalzeile identisch sind. Bei numerischen Attributen wird dabei eine Toleranz von einem Prozent berücksichtigt. Der Metrikwert entspricht dem Anteil der synthetischen Zeilen, die keinen solchen vollständigen Match aufweisen. Ein hoher Wert deutet darauf hin, dass der Generator in der Lage ist, neuartige Kombinationen von Attributen zu erzeugen, ohne bloße Kopien der Originaldaten zu produzieren.

Die Kombination dieser beiden Metriken erlaubt eine differenzierte Einschätzung der generativen Vielfalt. Sie beantwortet die zweite Forschungsfrage zur Reproduktion der Variabilität, indem sie sowohl die Abdeckung der ursprünglichen Wertebereiche als auch die Fähigkeit zur Generierung neuer Datenkombinationen betrachtet [5].

Datenschutz

Die Kategorie Datenschutz adressiert die dritte Forschungsfrage zur Vermeidung von Rückschlüssen auf Originaldaten. Zur Bewertung kommt die **Disclosure Protection** Metrik zum Einsatz, die simuliert, inwieweit sich aus den synthetischen Daten sensible Informationen rekonstruieren lassen. Dabei wird ein realistisches Angriffsszenario nachgebildet, bei dem ein Angreifer über bekannte Informationen wie Alter und Geschlecht verfügt und versucht, unbekannte sensible Attribute wie politische Zugehörigkeit zu erraten. Die Metrik vergleicht das resultierende Risiko mit einer Zufallsbaseline und liefert so eine Einschätzung der Schutzwirkung der synthetischen Daten [5].

Die Wahl dieser Metrik basiert auf ihrer inhaltlichen Nähe zur Forschungsfrage sowie auf ihrer technischen Verfügbarkeit innerhalb der SDMetrics Bibliothek. Sie bietet eine praktikable Möglichkeit, das Risiko ungewollter Rückschlüsse zu quantifizieren. Alternative Verfahren wie Membership Inference Tests oder Re-Identifikationsmetriken wurden nicht berücksichtigt, da sie andere Datenschutzkonzepte adressieren oder zusätzliche Anforderungen an die Datenbasis stellen, die im vorliegenden Setup nicht erfüllt sind .

Ergänzend zu dieser Einteilung wurde analysiert, ob sich bestimmte Konfigurationen von PrivBayes über alle Datensätze hinweg durch konsistent hohe Metrikwerte auszeichnen. Die Beantwortung dieser vierten Forschungsfrage erfolgte durch Aggregation der berechneten Metriken über alle Datensätze hinweg und durch die Identifikation leistungsstarker Konfigurationen.

Darüber hinaus wurde ein Vergleich der Modellfamilien PrivBayes, CTGAN und TVAE vorgenommen, um die Leistungsfähigkeit von PrivBayes im Vergleich zu modernen, auf neuronalen Netzen basierenden Generierungsverfahren einzuordnen. Die dazu verwendeten Vergleichsdaten wurden im Rahmen des Projekts INSIGHT durch die begleitende Forschungsgruppe bereitgestellt.

Durch die Kombination aus den im Evaluation Service verfügbaren SDMetrics Metriken, einer nachvollziehbaren inhaltlichen Strukturierung und der systematischen Verknüpfung mit den Forschungsfragen wird ein transparenter und methodisch fundierter Bewertungsansatz realisiert. Auch wenn die verwendeten Metriken nicht alle von Sarmin et al. [21], Du und Li [6] sowie Sella et al. [22] geforderten Dimensionen wie strukturelle oder semantische Aspekte vollständig abdecken, erlaubt die gewählte Auswahl eine praxisnahe und mehrdimensionale Einschätzung der Qualität synthetischer Daten, die über rein oberflächliche Ähnlichkeitsvergleiche hinausgeht.

3.4 Experimenteller Ablauf

Der Ablauf der Experimente gliedert sich in mehrere aufeinanderfolgende Schritte, die von der Datenaufbereitung bis zur Bewertung der synthetischen Datensätze reichen.

Zunächst wurden die fünf ausgewählten Originaldatensätze vorbereitet. Dazu zählten die Überprüfung auf fehlende Werte sowie bei zwei Datensätzen eine stichprobenartige Reduzierung der Datenmenge zur Sicherstellung der technischen Durchführbarkeit. Die weiteren Transformationen für die Anwendung von PrivBayes wurden durch die verwendete Bibliothek Synthcity automatisiert umgesetzt.

Anschließend erfolgte die Generierung der synthetischen Datensätze mit PrivBayes. Für jede Kombination der definierten Parameter k, Anzahl der Bins und ε wurden eigenständige synthetische Datensätze erzeugt. Insgesamt entstanden auf diese Weise 16 synthetische Varianten pro Originaldatensatz.

Die Qualität der synthetischen Daten wurde im Anschluss mithilfe des Evaluation Service der HAW Hamburg bewertet. Die Ergebnisse wurden entlang der zuvor beschriebenen Kategorien statistische Ähnlichkeit, Variabilität und Datenschutz ausgewertet. Zur systematischen Beantwortung der Forschungsfragen wurden die Metrikergebnisse sowohl innerhalb einzelner Datensätze als auch aggregiert über alle Datensätze hinweg analysiert.

Zur Einordnung der Ergebnisse wurden zusätzlich synthetische Datensätze der Modelle CTGAN und TVAE berücksichtigt, die von der begleitenden Forschungsgruppe zur Verfügung gestellt wurden.

4 Ergebnisse

In diesem Kapitel werden die fünf in der Einleitung (Kapitel 1) formulierten Forschungsfragen systematisch beantwortet. Grundlage der Auswertung sind die in Kapitel 3.3 beschriebenen Bewertungsmetriken, die gezielt entlang der Forschungsfragen ausgewählt und strukturiert wurden. Die Herleitung und Kategorisierung dieser Metriken ist in der Methodik (Kapitel 3) ausführlich erläutert, sodass an dieser Stelle der Fokus auf der Darstellung und Interpretation der Ergebnisse liegt.

Jede Forschungsfrage wird dabei einzeln betrachtet. Die jeweiligen Metrikergebnisse werden vorgestellt, kontextualisiert und abschließend bewertet, um eine differenzierte Einschätzung der Qualität und Nutzbarkeit der erzeugten synthetischen Daten zu ermöglichen.

Die synthetischen Datensätze von CTGAN und TVAE wurden im Rahmen des Forschungsprojekts INSIGHT durch die begleitende Forschungsgruppe erzeugt und dieser Arbeit zur Analyse zur Verfügung gestellt. Die Modelle basieren auf Implementierungen im Python-Framework Synthcity [18], wobei unterschiedliche Architekturen zum Einsatz kamen.

Einfluss der Binning-Stufe

Zur Einschätzung des Einflusses der Binning-Stufe auf die Qualität der synthetischen Daten wurden alle Konfigurationen mit 10 bzw. 20 Bins für den Adult-Datensatz gruppiert und die durchschnittlichen Werte der zentralen Metriken berechnet. Die jeweils acht Konfigurationen je Gruppe unterscheiden sich hinsichtlich der Parameter ε und k, sodass der Vergleich über verschiedene Datenschutzniveaus und Netzwerkkomplexitäten hinweg aggregiert erfolgt. Tabelle 4.1 zeigt die resultierenden Mittelwerte der beiden Gruppen.

Die Unterschiede zwischen den beiden Binning-Stufen fallen über alle Metriken hinweg gering aus. Lediglich bei den Column Pair Trends (0,025) und den Weighted Statistics (0,020) ist ein leichter Rückgang bei der Verwendung von 20 Bins zu beobachten. Die übrigen Metriken zeigen nahezu identische Werte oder geringfügige Verbesserungen bei 20 Bins. Da sich über alle betrachteten Konfigurationen hinweg kein systematischer Einfluss der Binning-Stufe auf die Datenqualität erkennen ließ, wurde im weiteren Verlauf auf eine vertiefte Analyse dieses Parameters verzichtet.

Tabelle 4.1: Durchschnittliche Metrikwerte für alle Konfigurationen mit 10 bzw. 20 Bins (Adult-Datensatz, jeweils 8 Varianten mit unterschiedlichen Kombinationen von ε und k).

Metrik	**Mittelwert 10 Bins**	**Mittelwert 20 Bins**	**Differenz**
Column Pair Trends	0,944	0,919	0,025
Weighted Statistics	0,978	0,958	0,020
Value Coverage	0,999	0,996	0,003
New Row Synthesis	0,962	0,964	−0,001
Disclosure Protection	0,862	0,866	−0,004

4.1 Wie gut erhält PrivBayes die statistischen Eigenschaften der Originaldaten?

Die erste Forschungsfrage untersucht, wie gut die statistischen Eigenschaften der Originaldaten in den synthetischen Daten erhalten bleibt. Die statistischen Eigenschaften umfassen dabei die Verteilungen einzelner Merkmale (univariate Verteilungen), die paarweisen Abhängigkeiten zwischen den Merkmalen (bivariate Verteilungen) sowie die Einhaltung der Wertebereiche. Eine verlässliche Reproduktion dieser Eigenschaften ist entscheidend für die Nutzbarkeit synthetischer Daten in analytischen Anwendungen.

Zur Beantwortung dieser Frage wurden drei Metriken verwendet: Die **Weighted Statistics** messen die Übereinstimmung der univariaten Verteilungen, die **Column Pair Trends** analysieren die bivariaten Abhängigkeiten, und die **Boundary Adherence** bewertet, ob die Wertebereiche der Originaldaten in den synthetischen Daten eingehalten werden. Abbildung 4.1 zeigt die Verteilung der betrachteten Metriken über alle synthetisierten Datensätze. Hervorzuheben ist, dass jede Metrik einen hohen bis perfekten Wert erreicht hat. Im Folgenden wird tiefer auf die Parametrisierung eingegangen.

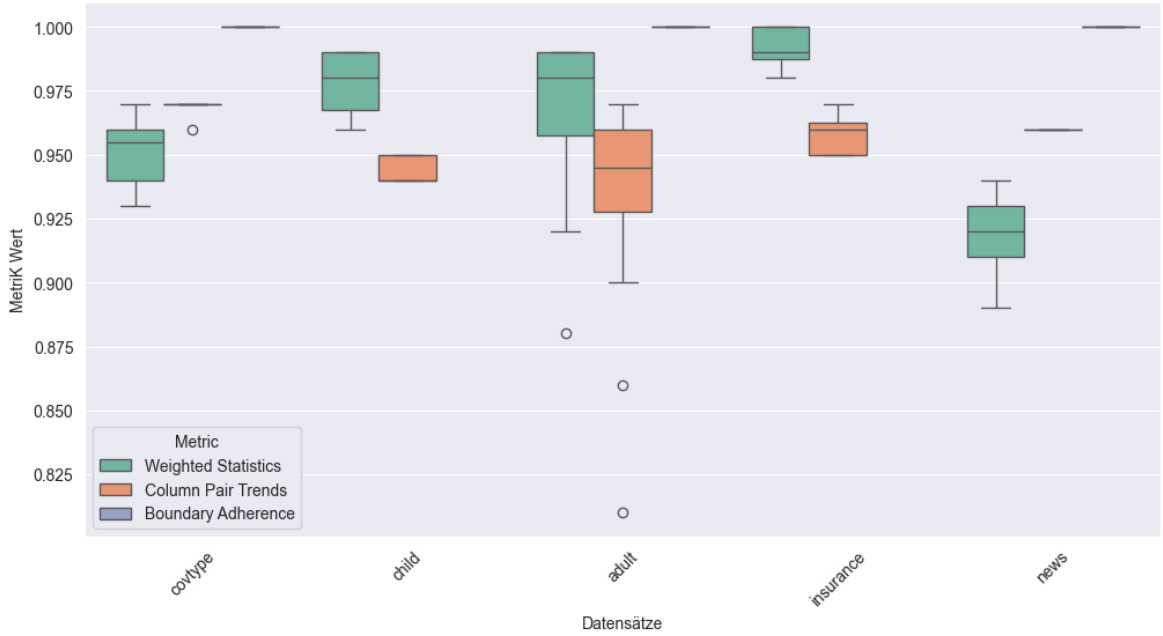


Abbildung 4.1: Verteilung der Weighted Statistics, Column Pair Trends und Boundary Adherence über alle synthetischen Datensätze

4.1.1 Analyse der Weighted Statistics

Die Ergebnisse der Weighted Statistics zeigen, wie stark die Parameter k und ϵ die Qualität der synthetischen Daten beeinflussen können. Abbildung 4.2 stellt die Verteilung der Weighted Statistics für die Datensätze in Abhängigkeit des gewählten k-Wertes dar. Die Metrik bewegt sich in einem hohen Bereich zwischen 0.88 und 1.0, was auf eine gute bis perfekte Übernahme der univariaten Verteilungen hinweist.

Bei den Datensätzen *Covtype* und *Child* führt die Wahl von $k = 2$ und $k = 3$ zu ähnlichen Ergebnissen, während bei *Adult*, *News* und *Insurance* deutliche Unterschiede sichtbar werden. Insbesondere verschlechtert sich bei diesen Datensätzen der Median bei $k = 3$ leicht, und die Ergebnisse werden variabler. Diese Variabilität zeigt sich in einer erhöhten Streuung, was darauf hindeutet, dass eine höhere Modellkomplexität ($k = 3$) nicht zwangsläufig eine bessere Annäherung an die Originalverteilung bewirkt. Vielmehr kann es je nach Datensatz dazu führen, dass zusätzliche Abhängigkeiten modelliert werden, die nicht in jedem Fall zur Verbesserung der Weighted Statistics beitragen.

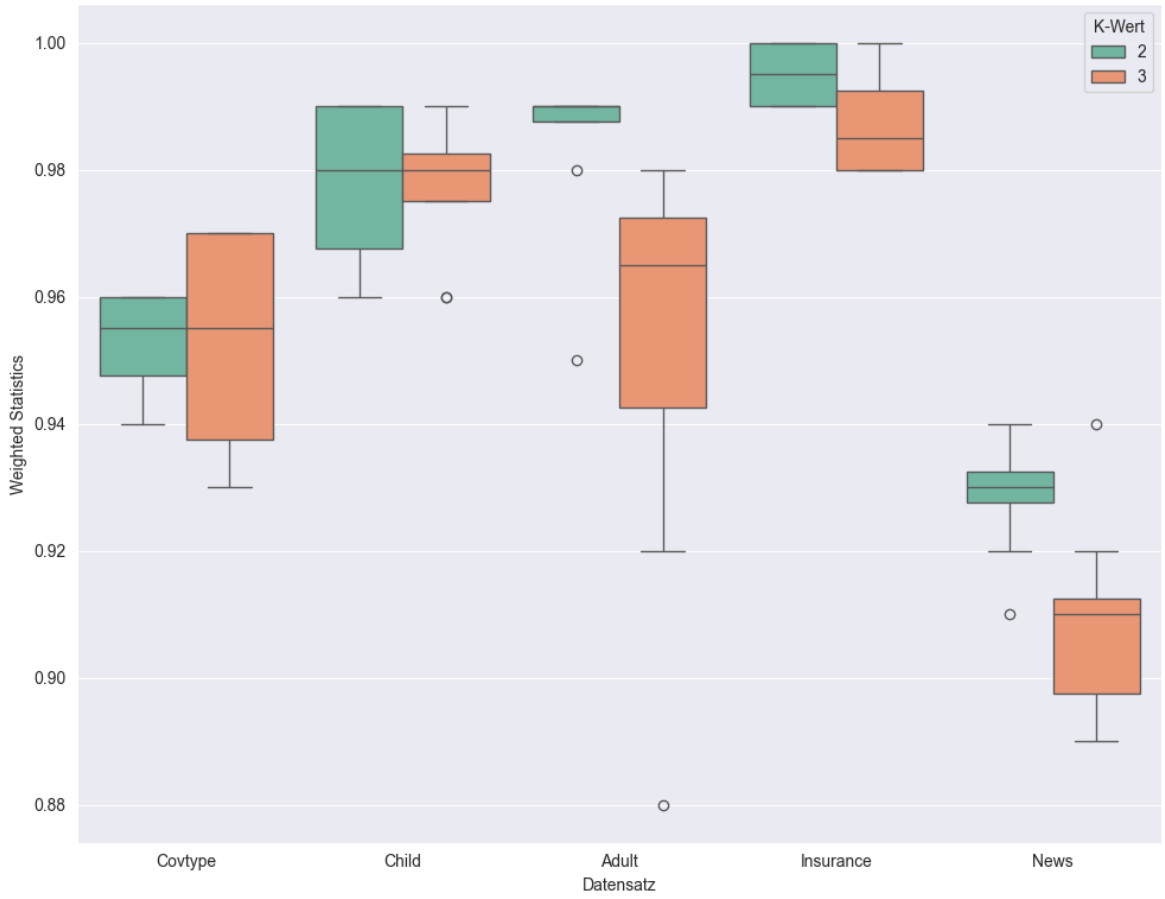


Abbildung 4.2: Weighted Statistics für alle Datensätze ($k = 2$ vs. $k = 3$)

Abbildung 4.3 zeigt die Weighted Statistics in Abhängigkeit vom gewählten ϵ-Wert. Ein kleines ϵ steht für ein höheres Datenschutzniveau, da mehr Rauschen in die Daten eingeführt wird. Während theoretisch höhere ϵ-Werte, also weniger Rauschen, eine stärkere Annäherung an die Originaldaten ermöglichen sollten, zeigt sich dies nicht bei allen Datensätzen. Die Datensätze *Covtype*, *Insurance* und *Child* weisen bei steigendem ϵ-Wert leicht verschlechterte Scores auf. Im Gegensatz dazu verhalten sich die Datensätze *News* und *Adult* erwartungsgemäß: Mit höheren ϵ-Werten verbessern sich die Weighted Statistics, wie an den höheren Medianwerten zu erkennen ist.

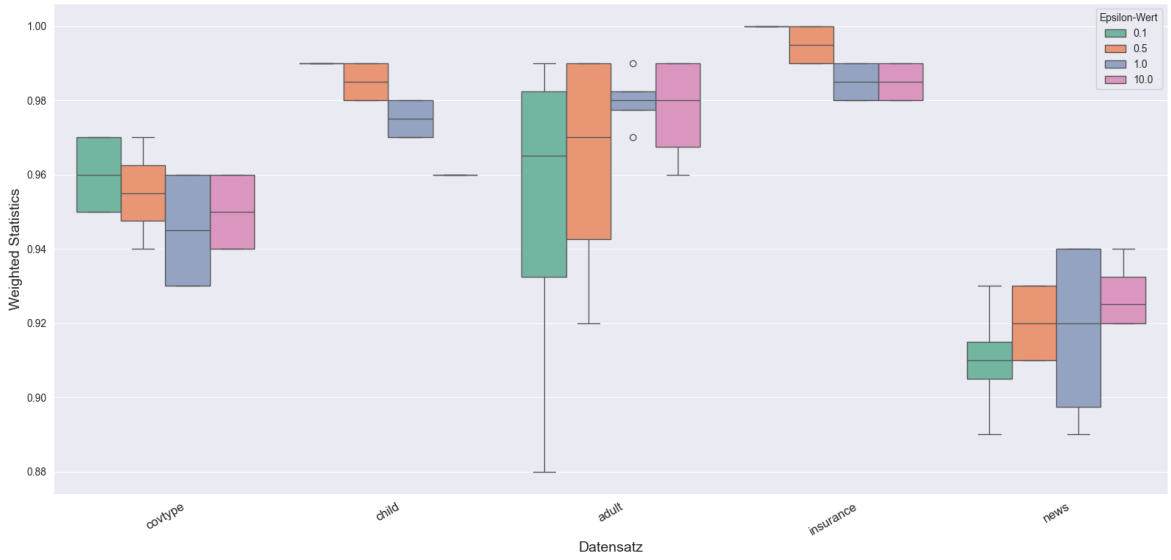
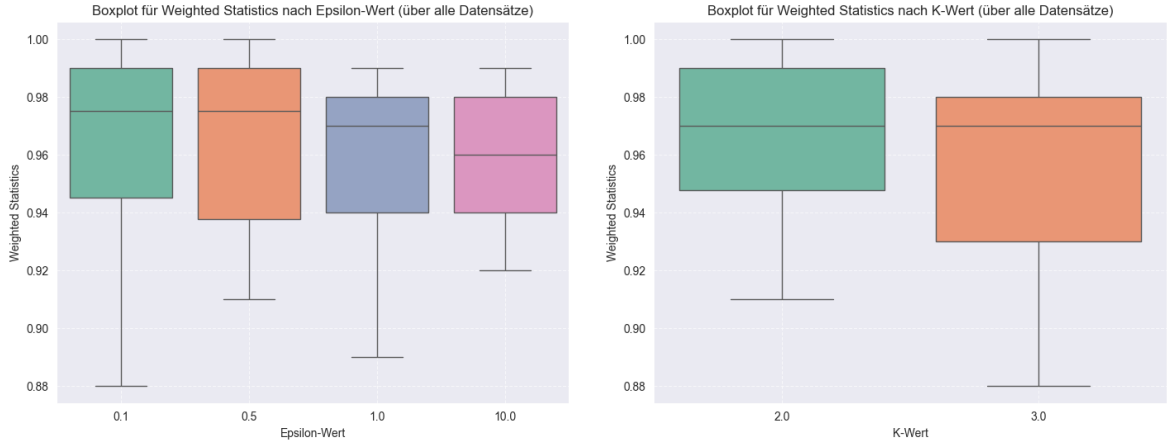


Abbildung 4.3: Weighted Statistics nach Datensatz und ϵ-Wert



(a) Weighted Statistics nach ϵ-Wert

(b) Weighted Statistics nach k-Wert

Abbildung 4.4: Vergleich der Weighted Statistics in Abhängigkeit von ϵ- und k-Wert gruppiert über alle Datensätze

Eine generelle Tendenz bei der Wahl von ϵ und k lässt sich in Abbildung 4.4 erkennen. Mit einer durchschnittlichen Weighted Statistics von 0.9675 erzielt $k = 2$ bessere Ergebnisse als $k = 3$ (0.9555). Dies deutet darauf hin, dass $k = 2$ eine vielversprechende Wahl für erste Analysen darstellt. Der Einfluss des Epsilon-Werts (ϵ) auf die Metrik ist vergleichsweise gering, da die durchschnittlichen Werte in einem engen Bereich zwischen 0.9600 und 0.9635 variieren. Dies

verdeutlicht, dass PrivBayes unabhängig von der Wahl des ϵ-Werts zuverlässig die univariaten Verteilungen repliziert und robust gegenüber unterschiedlichen Datenschutzniveaus agiert.

4.1.2 Analyse der Column Pair Trends

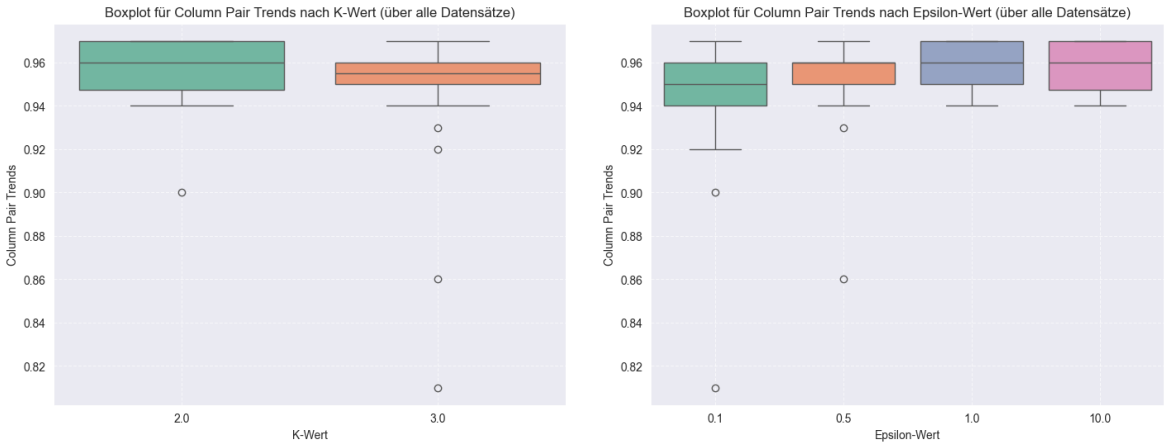
Die Column Pair Trends-Werte bewerten die Fähigkeit von PrivBayes, paarweise Abhängigkeiten zwischen Merkmalen zu erhalten. Alle Datensätze erzielen hohe Werte (siehe Abbildung 4.1), was darauf hindeutet, dass PrivBayes in der Lage ist, paarweise Abhängigkeiten gut zu reproduzieren. Besonders stabil sind die Ergebnisse bei den Datensätzen *Covtype*, *Child*, *Insurance* und *News*. Im Gegensatz dazu zeigt der *Adult*-Datensatz eine größere Streuung und niedrigere Werte. Dies deutet darauf hin, dass PrivBayes je nach Datensatz unterschiedlich gut funktioniert, wenn es darum geht, Beziehungen zwischen Merkmalen zu erhalten.

Abbildung A.2 verdeutlicht, dass der Parameterwert $k = 3$ lediglich beim *Child*-Datensatz positive Auswirkungen hat, da dort bessere Ergebnisse erzielt wurden. Im Gegensatz dazu führte $k = 3$ bei den Datensätzen *Adult* und *Insurance* zu schlechteren Ergebnissen. Bei den Datensätzen *Covtype* und *News* blieb der Einfluss von k hingegen minimal, da die Verteilungen nahezu identisch sind und keine nennenswerte Streuung aufweisen.

Diese geringe Streuung lässt sich auch bei der Analyse des ϵ-Wertes beobachten. Während die Datensätze *Adult* und *Insurance* mit steigendem ϵ (geringeres Datenschutzniveau) bessere Werte in der Metrik erzielen, bleiben die übrigen Datensätze stabil und zeigen kaum Variationen, unabhängig von ϵ (siehe Abbildung A.3).

Die in Abbildung 4.5a dargestellten Ergebnisse zeigen die Verteilung der Column Pair Trends-Werte über alle synthetischen Datensätze hinweg, gruppiert nach den Werten des Parameters k. Der Durchschnittswert für $k = 2$ liegt bei 0.95625, während für $k = 3$ ein geringerer Wert von 0.94900 erreicht wird. Dies verdeutlicht, dass $k = 2$ im Mittel leicht bessere Ergebnisse liefert. Zudem ist die Verteilung bei $k = 2$ kompakter, was auf eine größere Stabilität der Ergebnisse hinweist. Im Gegensatz dazu zeigt $k = 3$ eine größere Streuung mit einigen Ausreißern in den niedrigeren Bereichen.

Abbildung 4.5b zeigt die Verteilung der Column Pair Trends-Werte, gruppiert nach den Werten von ϵ ($\epsilon = 0.1$, $\epsilon = 0.5$, $\epsilon = 1.0$, $\epsilon = 10$). Die Durchschnittswerte steigen mit zunehmendem ϵ, wobei $\epsilon = 1.0$ und $\epsilon = 10$ den höchsten Mittelwert von 0.9580 aufweisen. Dies deutet darauf hin, dass größere ϵ-Werte, die einer geringeren Privacy-Beschränkung entsprechen, zu einer besseren Reproduktion der paarweisen Abhängigkeiten führen.



(a) Column Pair Trends nach k-Wert

(b) Column Pair Trends nach ϵ-Wert

Abbildung 4.5: Vergleich der Column Pair Trends in Abhängigkeit von ϵ- und k-Wert gruppiert über alle Datensätze

4.1.3 Analyse der Boundary Adherence

Bei den Datensätzen, für die Boundary Adherence anwendbar war, wurden durchweg perfekte Werte erzielt (siehe Abbildung 4.1). Dies zeigt, dass PrivBayes die Grenzwerte der Originaldaten zuverlässig einhält und keine Werte außerhalb des gelernten Wertebereichs synthetisiert. Dies ist eine direkte Folge der Bayesschen Netzwerkstruktur, die sicherstellt, dass nur beobachtete Werte generiert werden.

Insgesamt zeigt sich, dass PrivBayes in der Lage ist, die statistischen Eigenschaften der Originaldaten über verschiedene Metriken hinweg zuverlässig zu bewahren. Die Wahl von $k = 2$ und einem moderaten ϵ-Wert ($e \geq 1.0$) bietet ein gutes Gleichgewicht zwischen Stabilität und Genauigkeit.

4.2 Wie gut reproduziert PrivBayes die Variabilität der Originaldaten?

Die zweite Forschungsfrage untersucht, inwieweit die Variabilität der Originaldaten in den synthetischen Daten erhalten bleibt. Dabei steht im Fokus, ob die generierten Daten die Vielfalt der beobachteten Werte aus den Originaldaten widerspiegeln und ob PrivBayes in der Lage ist, neue, einzigartige Datenpunkte zu erzeugen, anstatt bestehende Werte lediglich zu replizieren.

Zur Beantwortung dieser Frage wurden zwei Metriken herangezogen: **Value Coverage** bewertet, ob alle in den Originaldaten vorkommenden Werte auch in den synthetischen Daten vertreten sind. Eine vollständige Abdeckung würde darauf hinweisen, dass keine Werte verloren gehen und die synthetischen Daten die gesamte Bandbreite der Originaldaten abbilden. Ergänzend dazu misst **New Row Synthesis** den Anteil der generierten Zeilen, die nicht direkt aus den Originaldaten übernommen wurden. Ein hoher Wert in dieser Metrik deutet darauf hin, dass PrivBayes nicht nur bestehende Werte nachbildet, sondern auch neue Datenpunkte generiert.

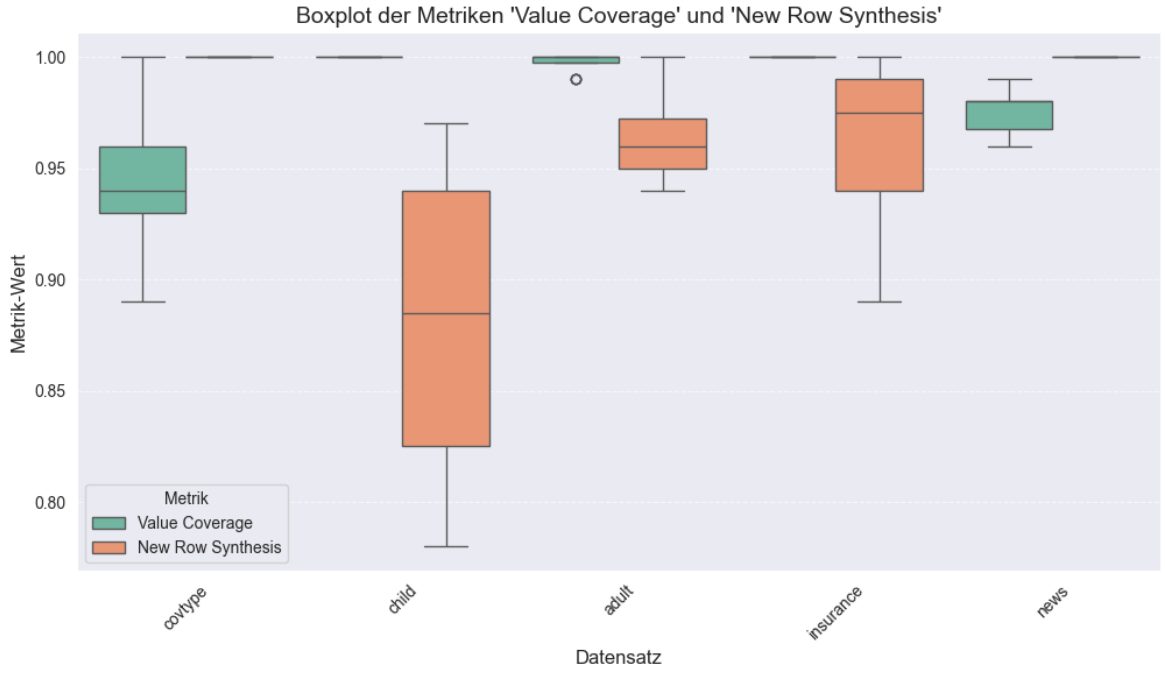


Abbildung 4.6: Boxplot der Value Coverage und New Row Synthesis über alle synthetischen Datensätze

Diese Metriken ermöglichen eine differenzierte Analyse darüber, ob PrivBayes die strukturelle Vielfalt der Originaldaten erhält und gleichzeitig eine gewisse Neuartigkeit in den synthetischen Daten sicherstellt. Abbildung 4.6 gibt einen ersten Überblick über die erzielten Metrik-Werte.

4.2.1 Analyse der Value Coverage

Die Value Coverage zeigt über alle Datensätze hinweg hohe bis nahezu perfekte Werte. Der *Covtype*-Datensatz weist dabei die größte Streuung auf, enthält jedoch mindestens eine Konfiguration mit einem perfekten Wert. Im Gegensatz dazu erreichte der *News*-Datensatz in keiner Konfiguration den Maximalwert von 1.0, zeigte jedoch insgesamt eine kompaktere Verteilung mit geringerer Streuung als *Covtype*.

Der Einfluss der gewählten Parameter ist insgesamt minimal ausgeprägt. Der Mittelwert über alle synthetisierten Datensätze, gruppiert nach k, erreichte für $k = 2$ einen Wert von 0.9842 und lag damit leicht über dem Wert für $k = 3$ mit 0.9832. Dies deutet darauf hin, dass eine geringere Modellkomplexität ausreicht, um die Vielfalt der beobachteten Werte in den synthetischen Daten zuverlässig zu erfassen. Eine ähnliche Tendenz zeigt sich für den Parameter ϵ, wobei die Konfigurationen mit $\epsilon = 1$ mit einem Durchschnitt von 0.9865 die besten Ergebnisse erzielten, dicht gefolgt von $\epsilon = 0.5$ mit 0.9850. Diese Unterschiede sind jedoch insgesamt gering, was darauf hindeutet, dass die Wahl dieser Parameter nur einen begrenzten Einfluss auf die Value Coverage hat.

Interessanterweise zeigt sich eine erhöhte Variabilität insbesondere bei den Datensätzen, bei denen die Anzahl der synthetisierten Zeilen reduziert wurde. Während der *News*-Datensatz von ursprünglich 5000 auf 500 Zeilen reduziert wurde, umfasst der *Covtype*-Datensatz nur 1000 statt der ursprünglichen 5000 Zeilen. Dies könnte darauf hindeuten, dass eine kleinere Stichprobe die Erfassung seltener Werte durch das Bayessche Netzwerk beeinflusst und sich somit auf die Value Coverage auswirkt. Allerdings zeigen die Unterschiede zwischen den Datensätzen, dass nicht allein die Zeilenanzahl ausschlaggebend ist, sondern auch die zugrunde liegende Datenstruktur eine Rolle spielt.

Eine weiterführende Analyse könnte untersuchen, ob die Reduktion der Zeilenanzahl zu einer Verringerung der Anzahl unterschiedlicher Merkmalsausprägungen geführt hat, die es dann zu reproduzieren gilt. Dies könnte den beobachteten Effekt weiter erklären. Allerdings erfordert eine solche Analyse einen detaillierten Vergleich der Verteilungen vor und nach der Reduktion für die Datensätze, was den Rahmen dieser Arbeit überschreiten würde.

Neben der allgemeinen Betrachtung der Value Coverage wurden auch Unterschiede zwischen kategorialen und kontinuierlichen Werten analysiert. Abbildung 4.7 zeigt für die synthetisierten *Adult*-Datensätze, dass die Coverage für kontinuierliche Werte (*Range Coverage*) im Vergleich zu kategorialen Werten (*Category Coverage*) leicht variieren kann.

Die Ergebnisse legen nahe, dass beide Attributtypen durch den synthetischen Prozess gut abgedeckt werden, wobei die Werte für kontinuierliche Attribute in manchen Fällen etwas variabler sind. Dies könnte darauf zurückzuführen sein, dass kontinuierliche Werte grundsätzlich schwieriger exakt zu replizieren sind als kategoriale Werte, da sie theoretisch unendlich viele mögliche Ausprägungen haben. Während kategoriale Merkmale diskrete Werte annehmen und damit einfacher vollständig abgedeckt werden können, kann die Range Coverage bei kontinuierlichen Werten geringfügig niedriger ausfallen.

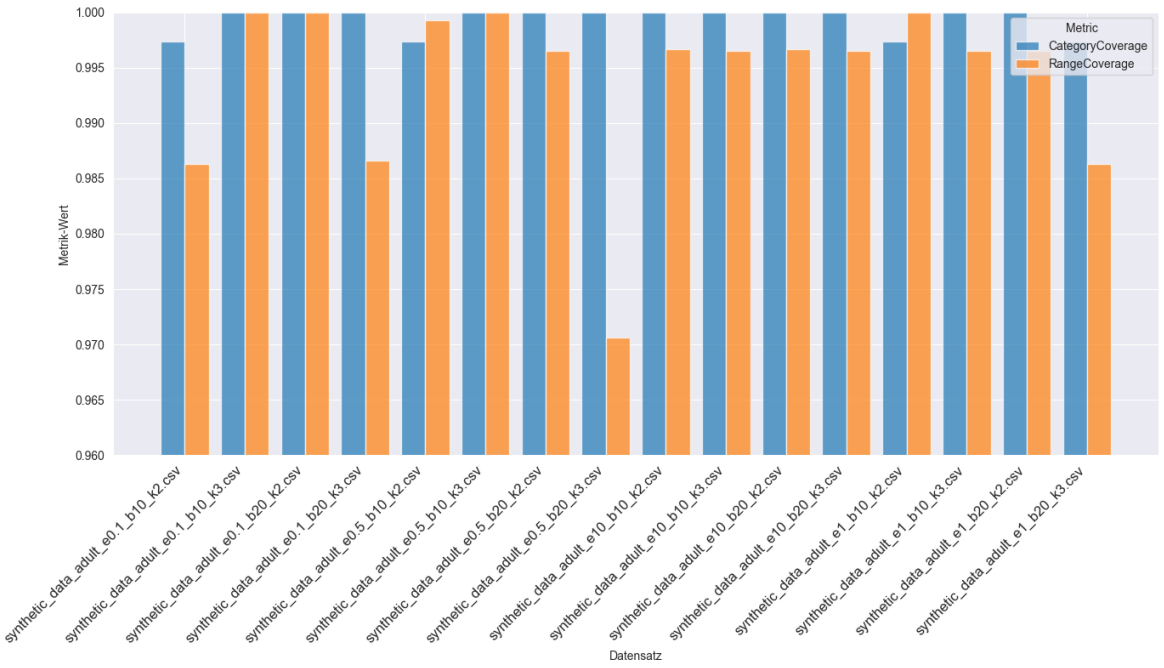


Abbildung 4.7: Unterschiede in der Value Coverage zwischen kategorialen (*Category Coverage*) und kontinuierlichen (*Range Coverage*) Werten für unterschiedliche Parameterkonfigurationen des Adult-Datensatzes

Es stellt sich zudem die Frage, ob die beobachteten Unterschiede tatsächlich auf die gewählten Parameter (ϵ und k) zurückzuführen sind oder ob sie vielmehr eine grundsätzliche Eigenschaft der Metrik widerspiegeln. Da kontinuierliche Werte eine unendliche Skala annehmen können, ist es unwahrscheinlicher, dass alle Wertebereiche exakt abgedeckt werden, was zu einer natürlich geringeren Range Coverage führt. Daher könnte es sich hierbei um eine systematische Tendenz handeln, dass die Coverage für kontinuierliche Werte häufig etwas niedriger ausfällt als für kategoriale Werte, unabhängig von den gewählten Parametern. Dennoch bleibt die allgemeine Abdeckung für beide Wertetypen auf einem hohen Niveau (0.96), was die Robustheit der synthetischen Datenerzeugung für verschiedene Attributtypen unterstreicht.

4.2.2 Analyse der New Row Synthesis

Die Ergebnisse der New Row Synthesis (Abbildung 4.6) zeigen, dass für die Datensätze *Covtype* und *News* weder der Parameter k noch ϵ einen entscheidenden Einfluss hatten. Über alle Konfigurationen hinweg erzielten diese Datensätze perfekte Werte.

Ein möglicher Einflussfaktor für diese hohen Werte könnte die reduzierte Anzahl der Zeilen sein. Vor der Generierung der synthetischen Daten wurden die Originaldatensätze gekürzt, und die synthetischen Datensätze wurden mit derselben Anzahl an Zeilen erzeugt wie die reduzierten Originaldatensätze. Eine geringere Anzahl an generierten Zeilen könnte die Wahrscheinlichkeit verringern, dass identische Einträge aus den Originaldaten in den synthetischen Daten wieder auftauchen. Dies würde höhere Werte in der New Row Synthesis begünstigen.

Allerdings zeigen die Unterschiede zwischen *Adult*, *Child* und *Insurance*, dass neben der Zeilenanzahl auch andere Faktoren, wie die zugrunde liegende Datenstruktur, eine Rolle spielen.

Der Einfluss des Parameters k auf die New Row Synthesis ist minimal. Der Mittelwert über alle Datensätze gruppiert nach k ergab 0.9625 für $k = 2$ und 0.9600 für $k = 3$, was nur einen geringen Unterschied darstellt.

Anders verhält es sich mit dem Einfluss von ϵ. Abbildung 4.8 zeigt deutlich, dass mit steigendem ϵ-Wert (abnehmendem Datenschutz) die Anzahl neuartiger synthetischer Zeilen abnimmt. Dies ist darauf zurückzuführen, dass mit höherem ϵ weniger Rauschen hinzugefügt wird, wodurch das Modell stärker an den Originaldaten ausgerichtet bleibt und existierende Muster exakter reproduziert werden.

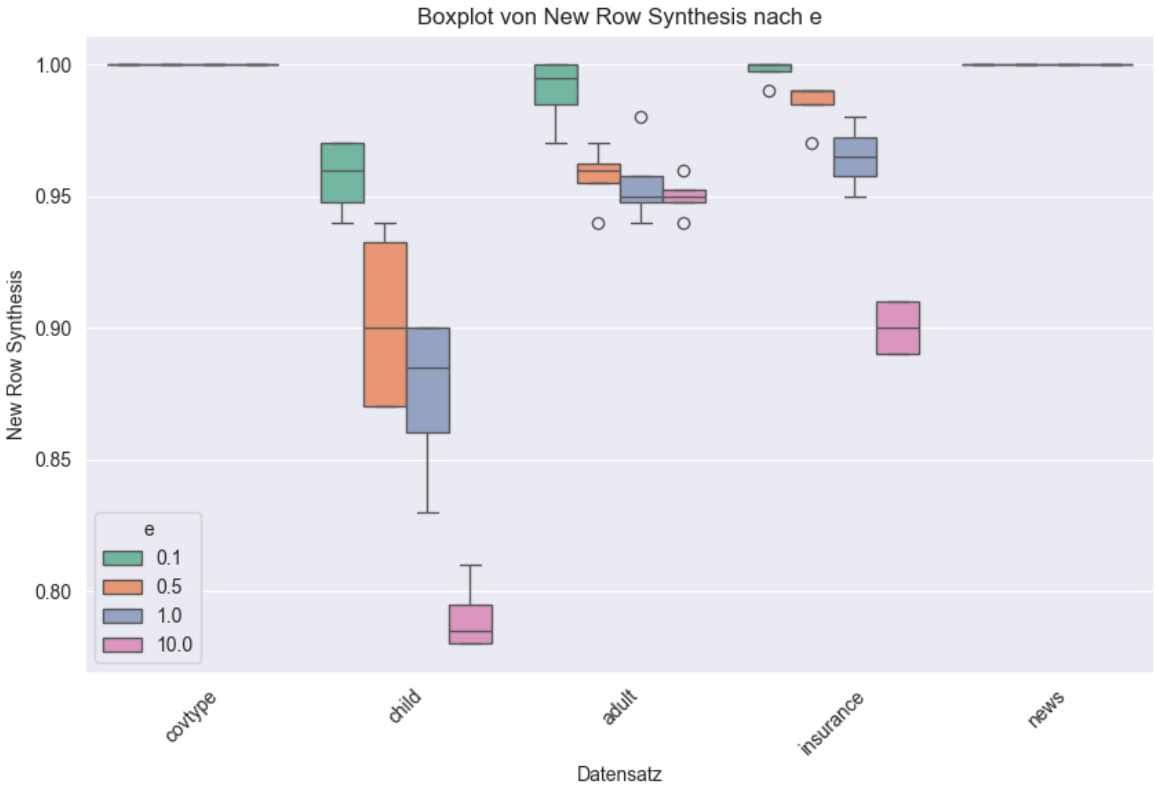


Abbildung 4.8: New Row Synthesis gruppiert nach Datensatz und ϵ-Wert

Die New Row Synthesis-Metrik bietet dabei nicht nur eine Aussage über die Variabilität der synthetischen Daten, sondern stellt zugleich eine einfache Privacy-Metrik dar. Sie liefert einen ersten Hinweis darauf, wie stark sich die generierten Daten von den Originaldaten unterscheiden. Ein niedriger Wert würde darauf hindeuten, dass viele der synthetischen Zeilen mit den Originaldaten übereinstimmen, was datenschutztechnisch problematisch sein könnte. Ein hoher Wert hingegen spricht dafür, dass neue Datenpunkte generiert wurden, was für den Schutz sensibler Informationen vorteilhaft sein kann.

Da Privatsphäre ein zentrales Kriterium in der Generierung synthetischer Daten darstellt, wird mit der folgenden Frage detaillierter untersucht, inwieweit PrivBayes den Schutz sensibler Informationen gewährleistet. Während New Row Synthesis eine grundlegende Abschätzung

erlaubt, wird mit der Disclosure Protection-Metrik eine detailliertere Analyse vorgenommen, um die mögliche Rückführbarkeit synthetischer Daten auf die Originaldaten zu bewerten.

4.3 Inwieweit gewährleistet PrivBayes den Schutz sensibler Informationen?

Der Schutz sensibler Informationen ist eine zentrale Anforderung bei der Generierung synthetischer Daten, insbesondere wenn diese für datenschutzkritische Anwendungen genutzt oder weitergegeben werden sollen. Dies ist besonders relevant in Bereichen, in denen Originaldaten aufgrund gesetzlicher Vorgaben oder ethischer Richtlinien nicht geteilt werden dürfen. Synthetische Daten ermöglichen Datenanalysen und Kooperationen, ohne sensible Informationen preiszugeben mit der Voraussetzung, dass der Datenschutz gewährleistet ist.

Zur Beurteilung der Datenschutzqualität wurde die **Disclosure Protection**-Metrik verwendet. Sie bewertet, wie gut PrivBayes verhindert, dass sensible Informationen aus den synthetischen Daten abgeleitet werden können. Die Metrik simuliert ein realistisches Angriffsszenario, bei dem ein Angreifer bekannte Informationen (zum Beispiel Alter und Geschlecht) nutzt, um unbekannte sensible Attribute (zum Beispiel politische Zugehörigkeit) zu erraten.

Ein hoher Wert (nahe 1.0) weist darauf hin, dass die synthetischen Daten ein ähnlich geringes Risiko wie zufällig generierte Daten bergen und keine zusätzlichen Informationen offenlegen.

In diesem Abschnitt wird analysiert, wie gut PrivBayes den Schutz sensibler Informationen gewährleistet und welche Auswirkungen verschiedene Parameterkonfigurationen auf die Disclosure Protection haben.

In Abbildung 4.9, wird die Metrik gruppiert nach Datensatz gezeigt. Lediglich *Covtype* und *News* enthielten Konfigurationen, die einen perfekten Wert (1.0) erzielten. Während *News* insgesamt sehr hohe und konsistente Werte aufweist, zeigt *Covtype* eine starke Streuung mit Werten, die bis auf etwa 0.3 fallen. Auch der *Insurance*-Datensatz weist eine deutliche Streuung auf, mit Werten zwischen ca. 0.5 und 0.87.

Im Vergleich dazu sind die Datensätze *Child*, *Adult* und *News* kompakter, was auf eine geringere Varianz in der Metrik hindeutet. Besonders *News* zeigt nicht nur die höchsten Medianwerte, sondern auch die konsistentesten Ergebnisse mit einer sehr kleinen Spannweite. In den Datensätzen *Adult* und *News* sind zudem einzelne Ausreißer zu erkennen, die auf Konfigurationen mit abweichenden Schutzwerten hinweisen.

Die beobachteten Unterschiede in der Streuung könnten darauf hindeuten, dass die gewählten Konfigurationen einen Einfluss auf die Höhe der Disclosure Protection haben. Insbesondere bei den Datensätzen *Covtype* und *Insurance* zeigt sich eine größere Varianz, was auf sensitivere Reaktionen auf unterschiedliche Einstellungen schließen lässt.

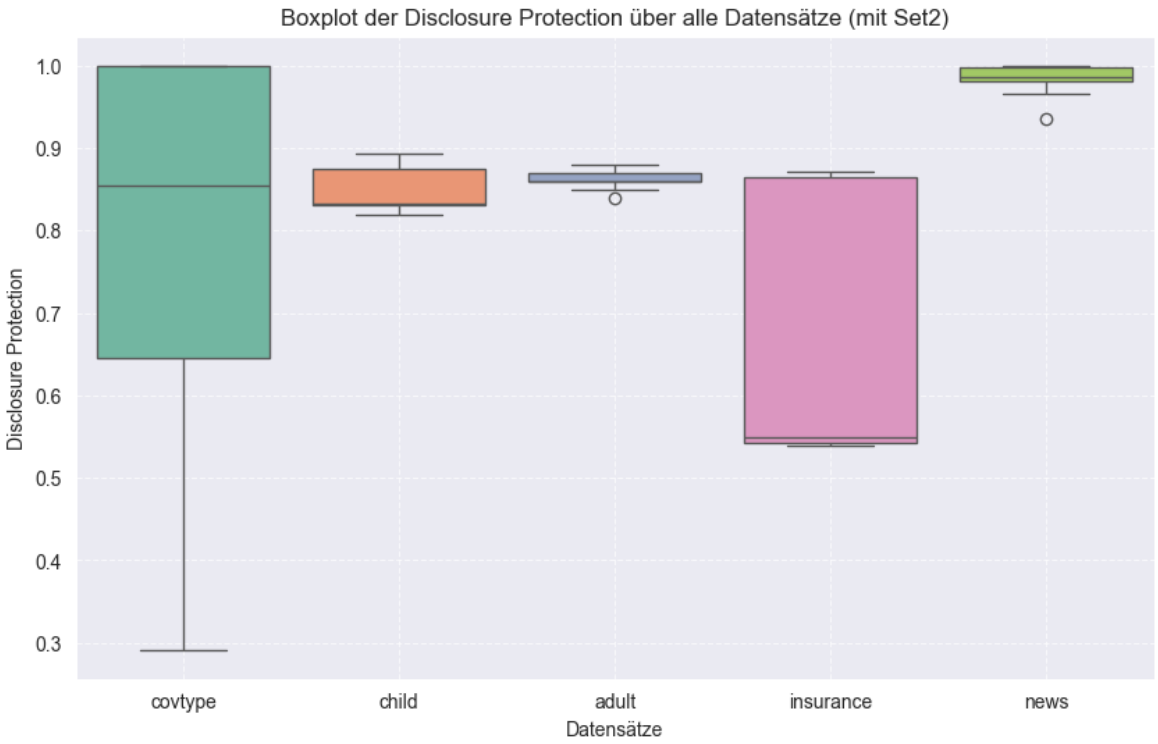
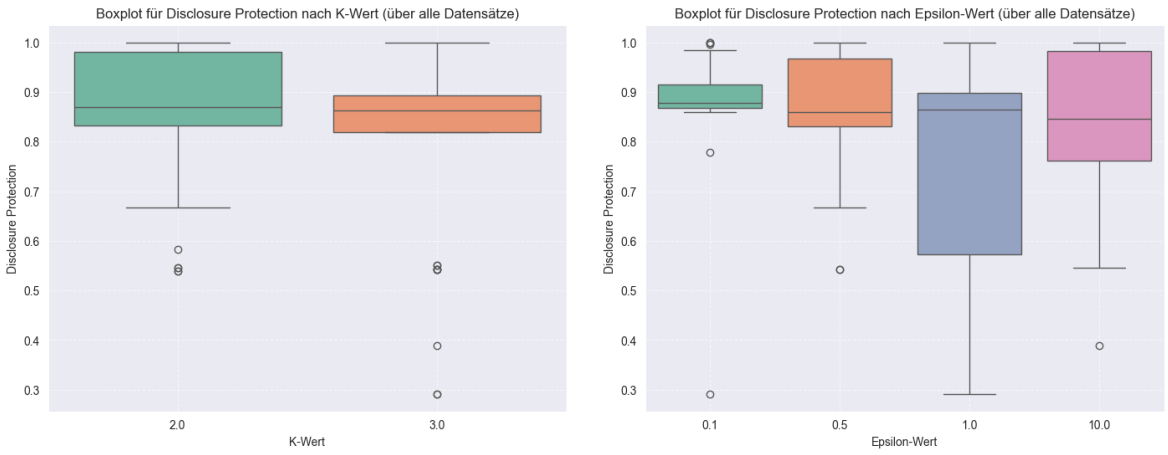


Abbildung 4.9: Verteilung der Disclosure Protection über alle synthetischen Datensätze



(a) Disclosure Protection nach k-Wert

(b) Disclosure Protection nach ϵ-Wert

Abbildung 4.10: Vergleich der Disclosure Protection in Abhängigkeit vom k- und ϵ-Wert über alle Datensätze

In Abbildung 4.10a ist der Einfluss des k-Werts auf die Disclosure Protection über alle Datensätze hinweg dargestellt. Der Vergleich zwischen den k-Werten 2 und 3 zeigt deutliche Unterschiede in den Schutzwerten. Generell ist zu erkennen, dass der Median der Metrik bei $k = 2$ höher liegt als bei $k = 3$, was darauf hindeutet, dass ein niedrigerer k-Wert tendenziell besseren Schutz bietet. Gleichzeitig ist die Streuung der Werte bei $k = 2$ größer. Während ein Großteil der Werte im oberen Bereich zwischen 0.85 und 1.0 liegt, sind auch einige niedrigere Ausreißer (bis ca. 0.55) erkennbar. Dies deutet darauf hin, dass $k = 2$ zwar im Durchschnitt einen höheren Schutz gewährleistet, die Ergebnisse jedoch variabler sind. Der durchschnittliche Disclosure Protection-Wert für $k = 2$ liegt bei 0.855, was den beobachteten Trend bestätigt.

Im Gegensatz dazu zeigt der k-Wert 3 insgesamt kompaktere Werte mit weniger Streuung. Die Schutzwerte sind hier stabiler, der Median jedoch niedriger als bei $k = 2$. Auch bei $k = 3$ sind Ausreißer vorhanden, die in einzelnen Fällen deutlich niedrigere Werte (bis ca. 0.29) erreichen. Der durchschnittliche Disclosure Protection-Wert für $k = 3$ beträgt 0.800, was den Rückgang der Schutzqualität bei einem höheren k-Wert unterstreicht. Diese Beobachtungen deuten darauf hin, dass ein höherer k-Wert konsistentere, aber insgesamt schwächere Schutzwerte liefert.

Zusammenfassend lässt sich sagen, dass ein niedrigerer k-Wert ($k = 2$) im Durchschnitt zu besseren Schutzwerten führt (0.855), jedoch mit einer höheren Varianz einhergeht. Ein höherer k-Wert ($k = 3$) hingegen bietet stabilere, aber tendenziell schwächere Schutzwerte (0.800). Abhängig vom Anwendungsfall könnte daher ein niedrigerer k-Wert bevorzugt werden, wenn ein maximaler Schutz angestrebt wird, während ein höherer k-Wert geeigneter sein könnte, wenn Konsistenz wichtiger ist als der absolute Schutzwert.

In Abbildung 4.10b ist der Einfluss des ϵ-Werts auf die Disclosure Protection über alle Datensätze hinweg dargestellt. Der Vergleich der unterschiedlichen ϵ-Werte zeigt deutliche Unterschiede in den Schutzwerten sowie in der Streuung der Ergebnisse.

Bei einem niedrigen ϵ-Wert von 0.1 sind die Schutzwerte insgesamt hoch und stabil. Der Median liegt im oberen Bereich (ca. 0.9) und die Streuung ist gering, was auf einen starken Schutz mit konsistenten Ergebnissen hindeutet. Dies spiegelt sich auch im durchschnittlichen Disclosure Protection-Wert von 0.871 wider, dem höchsten Durchschnitt aller getesteten ϵ-Werte. Einzelne Ausreißer zeigen jedoch, dass es in wenigen Fällen zu deutlich niedrigeren Schutzwerten kommen kann.

Mit zunehmendem ϵ-Wert verändert sich das Bild: Bei 0.5 bleibt der Median hoch, die Streuung nimmt jedoch leicht zu, was auf eine größere Varianz der Schutzwerte hindeutet. Der Durchschnittswert liegt hier bei 0.849, was zwar etwas unter dem Wert für $\epsilon = 0.1$ liegt, aber immer noch einen insgesamt hohen Schutz darstellt.

Besonders auffällig ist der ϵ-Wert 1.0, bei dem die Schutzwerte deutlich streuen. Hier reicht die Spanne von hohen Werten bis hinunter zu ca. 0.3, was auf eine erhöhte Unsicherheit und Variabilität der Schutzmechanismen hinweist. Der Durchschnittswert sinkt in diesem Fall

auf 0.783, den niedrigsten Schutzwert bei allen getesteten ϵ-Werte, was den Rückgang der Schutzqualität bei mittleren ϵ-Werten verdeutlicht.

Interessanterweise stabilisieren sich die Schutzwerte wieder bei einem hohen ϵ-Wert von 10.0. Der Median bleibt relativ hoch, während die Streuung im Vergleich zu $\epsilon = 1.0$ wieder abnimmt. Der Durchschnittswert steigt wieder leicht auf 0.808, was darauf hindeutet, dass bei sehr großen ϵ-Werten bestimmte Effekte, die die Streuung erhöhen, abgeschwächt werden.

Insgesamt zeigt der Vergleich, dass kleinere ϵ-Werte (insbesondere $\epsilon = 0.1$) zu höheren und stabileren Schutzwerten führen. Die durchschnittliche Disclosure Protection ist bei $\epsilon = 0.1$ mit 0.871 am höchsten und sinkt mit steigendem ϵ-Wert auf 0.783 bei $\epsilon = 1.0$ ab. Bei sehr hohen ϵ-Werten (10.0) verbessert sich der Durchschnittswert wieder leicht auf 0.808. Dies verdeutlicht den Einfluss des ϵ-Werts auf die Balance zwischen Datenschutz und Datenverfügbarkeit: Niedrigere ϵ-Werte fördern stärkeren Schutz auf Kosten der Genauigkeit, während höhere Werte mehr Genauigkeit ermöglichen, jedoch zu schwankenden Schutzwerten führen können.

4.4 Gibt es eine Konfiguration, die über alle Datensätze hinweg gute Ergebnisse erzielt?

Um die Forschungsfrage zu beantworten wurde untersucht, ob es Konfigurationen gibt, die in allen zuvor definierten Metriken konsistent gute Ergebnisse erzielen. Eine Konfiguration wurde als „gut“ eingestuft, wenn sie in jeder Metrik einen Wert von mindestens 0.9 erreicht.

Zunächst wurden die Daten hinsichtlich ihrer Vollständigkeit geprüft. Dabei zeigte sich erneut, dass die Metrik Boundary Adherence nicht für alle Konfigurationen berechnet werden konnte, da sie für bestimmte Datensätze nicht anwendbar war. Diese fehlenden Werte wurden bei der Berechnung der Durchschnittswerte ausgeschlossen.

Anschließend wurden die Daten pro Konfiguration aggregiert. Für jede Kombination aus k, ϵ und Anzahl der Bins wurden die bereits ermittelten Metrikergebnisse aus den verschiedenen Datensätzen zusammengeführt. Anschließend wurde für jede Metrik der Durchschnitt berechnet, indem die zusammengeführten Werte durch die Anzahl der verfügbaren Metriken für die jeweilige Konfiguration geteilt wurden.

Dieses Vorgehen ermöglicht eine vergleichende Bewertung der Konfigurationen über verschiedene Datensätze hinweg, um Muster in der Gesamtleistung zu identifizieren. Allerdings ist zu beachten, dass durch die Aggregation der Metrikergebnisse Unterschiede in der Performance einzelner Datensätze nicht mehr sichtbar sind. Der Fokus liegt hier bewusst auf der Identifikation robuster Konfigurationen, die sich unabhängig von datensatzspezifischen Eigenschaften konsistent bewähren. Die detaillierten Metrikergebnisse pro Datensatz wurden zuvor einzeln betrachtet und visualisiert.

Beispiel: Für die Konfiguration $k = 2$, $\epsilon = 0.1$ und 10 *bins* wurde für jede einzelne Metrik (z. B. *Boundary Adherence*, *Weighted Statistics*, *Value Coverage*) die bereits berechneten Werte aus den verfügbaren Datensätzen summiert und durch die Anzahl der vorliegenden Werte für die Metrik geteilt.

Im Anschluss daran wurden alle Konfigurationen dahingehend geprüft, ob sie in sämtlichen Metriken den festgelegten Schwellenwert von 0.9 überschreiten. Ziel war es, Konfigurationen zu identifizieren, die eine konsistent hohe Leistung über alle Metriken hinweg aufweisen und nicht nur in einzelnen Metriken gute Ergebnisse erzielen.

Die Analyse ergab zwei Konfigurationen, die diese Anforderungen erfüllten. Beide wiesen denselben Wert für $k = 2$ sowie eine Anzahl von 10 *bins* auf, während der ϵ-Wert zwischen 0.1 und 0.5 variierte.

Die Tabellen 4.2 und 4.3 zeigen, dass beide Konfigurationen in allen Metriken Werte oberhalb des festgelegten Schwellenwerts von 0.9 erreichen. Darüber hinaus zeigen alle Metriken konsistent hohe Werte, was auf insgesamt stabile und leistungsstarke Konfigurationen hinweist.

k	ϵ	*bins*	**Column Pair Trends**	**Weighted Statistics**	**Boundary Adherence**
2	0.1	10	0.952	0.972	1.000
2	0.5	10	0.958	0.974	1.000

Tabelle 4.2: Durchschnittswerte der identifizierten „guten“ Konfigurationen – Teil 1.

k	ϵ	*bins*	**Value Coverage**	**New Row Synthesis**	**Disclosure Protection**
2	0.1	10	0.978	0.988	0.925
2	0.5	10	0.984	0.978	0.910

Tabelle 4.3: Durchschnittswerte der identifizierten „guten“ Konfigurationen – Teil 2.

Diese Ergebnisse verdeutlichen, dass insbesondere Konfigurationen mit einem niedrigen k-Wert und einer geringeren Anzahl an *bins* in dieser aggregierten Betrachtung tendenziell bessere Resultate liefern. Zusätzlich legen sie nahe, dass die Motivation hinter der Synthese klar definiert sein sollte. Beide Konfigurationen unterscheiden sich in ihrer Wahl des ϵ-Werts.

Falls das Hauptziel darin liegt, der statistischen Struktur (Weighted Statistics, Column Pair Trends, Boundary Adherence) mehr Gewicht zu geben, wäre die Konfiguration mit $\epsilon = 0.5$ zu bevorzugen, da sie leicht bessere Ergebnisse erzielt. Falls hingegen der Datenschutz (Disclosure Protection) im Vordergrund steht, wäre $\epsilon = 0.1$ die bessere Wahl.

Zwar sind die Unterschiede zwischen den beiden Konfigurationen in dieser Analyse gering, doch könnten sie bei Datensätzen mit anderer Struktur deutlicher ausfallen. Dies könnte dazu führen, dass ein Akteur in diesem Fall eine bewusste Entscheidung zwischen statistischer Genauigkeit und Datenschutz treffen muss.

4.5 Wie unterscheiden sich die Ergebnisse von PrivBayes, CTGAN und TVAE?

Um die Frage zu beantworten wurde ein Vergleich der Modelle durchgeführt. Die Modelle wurden in unterschiedlichen Konfigurationen mit den gleichen Originaldaten trainiert und anschließend wurden daraus Daten synthetisiert.

Die Metrikwerte für die verschiedenen Modellkonfigurationen wurden im Rahmen von Experimenten der Forschungsgruppe ermittelt. Diese Ergebnisse dienten als Grundlage für eine weiterführende Analyse, um die leistungsfähigsten Modellkonfigurationen zu identifizieren.

Wie bereits im vorherigen Abschnitt 4.4 beschrieben, wurden für PrivBayes die gleichen Konfigurationen über alle Datensätze hinweg aggregiert. Dabei wurde für jede Konfiguration jede einzelne Metrik, die in verschiedenen Datensätzen berechnet wurde, summiert und anschließend durch die Anzahl der Datensätze geteilt, in denen diese Konfiguration verwendet wurde. Auf diese Weise wurde ein Durchschnittswert für jede Metrik innerhalb einer Konfiguration ermittelt. Basierend auf diesen aggregierten Werten konnten zwei Konfigurationen identifiziert werden, die in allen Metriken Werte über 0.9 erreichten und somit als „gut“ eingestuft wurden.

Das gleiche Verfahren wurde für CTGAN und TVAE angewandt. Hierbei zeigte sich jedoch, dass keine der getesteten Konfigurationen in allen Metriken den Schwellenwert von 0.9 erreichte. Um dennoch eine fundierte Auswahl zu treffen, wurde auf Grundlage der bereits aggregierten Metrik-Durchschnittswerte innerhalb jeder Konfiguration ein zusätzlicher Schritt durchgeführt: die Berechnung des Durchschnitts über alle Metriken innerhalb einer Konfiguration.

Durch diese zusätzliche Aggregation konnte eine Gesamtbewertung der Konfigurationen erstellt werden, die eine objektive Grundlage für den Vergleich bietet. Anschließend wurden die zwei Konfigurationen mit den höchsten Durchschnittswerten für CTGAN und TVAE als die leistungsfähigsten Modelle ausgewählt, da sie über alle Metriken hinweg die besten Gesamtwerte erzielten.

Dieses zweistufige Verfahren gewährleistet, dass auch für CTGAN und TVAE die leistungsstärksten Konfigurationen bestimmt werden konnten, selbst wenn keine einzelne Konfiguration in allen Metriken über 0.9 lag. Durch diese Methodik wird sichergestellt, dass eine objektive und nachvollziehbare Auswahl der besten Modellkonfigurationen erfolgt.

Die in Abbildung 4.11 dargestellten Balkendiagramme zeigen die Metrikwerte für die verschiedenen Modellkonfigurationen von TVAE, CTGAN und PrivBayes. Durch diese Visualisierung lassen sich die Unterschiede zwischen den Modellfamilien besonders deutlich erkennen.

Auffällig ist, dass PrivBayes in nahezu allen Metriken konstant hohe Werte erzielt. Besonders in den Bereichen Column Pair Trends, New Row Synthesis, Value Coverage und Weighted Statistics erreichen die beiden PrivBayes-Konfigurationen durchweg Werte über 0.95, während

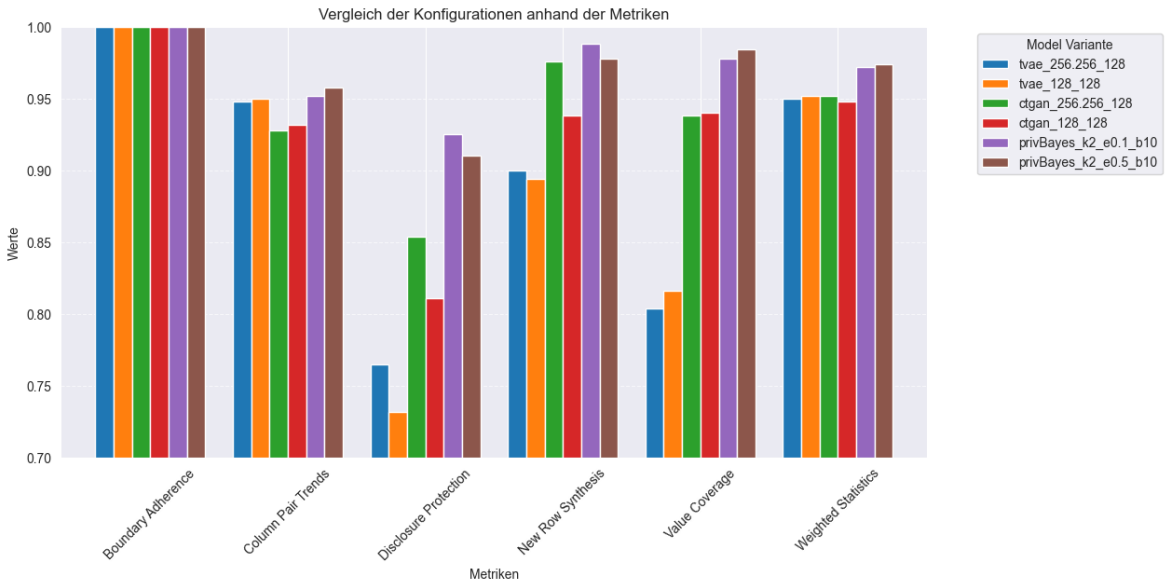


Abbildung 4.11: Vergleich der Metriken in den besten identifizierten Konfigurationen

die anderen Modellfamilien stärkere Schwankungen aufweisen. Dies deutet darauf hin, dass PrivBayes eine konsistente und stabile Leistung erbringt, insbesondere im Hinblick auf die Qualität und Vielfalt der synthetisierten Daten.

Im Vergleich dazu zeigen CTGAN- und TVAE-Modelle größere Variationen in ihrer Performance. Insbesondere in der Metrik Disclosure Protection fallen mehrere CTGAN- und TVAE-Konfigurationen deutlich ab, wobei einige Werte unter 0.8 liegen.

Ein zentraler Unterschied liegt darin, dass PrivBayes explizit auf den Schutz der Privatsphäre ausgelegt ist, während CTGAN und TVAE primär darauf abzielen, die Originaldaten möglichst realistisch nachzubilden. Dies erklärt, warum die Privacy-Metrik bei PrivBayes durchgehend höhere Werte erzielt, während die leistungsfähigeren generativen Modelle auf Basis neuronaler Netze hier Schwächen aufweisen.

Auch im Bereich New Row Synthesis zeigen sich klare Unterschiede: Während PrivBayes durchweg hohe Werte erzielt, schwanken die Ergebnisse bei CTGAN und TVAE stärker. Dies deutet darauf hin, dass PrivBayes konsistenter neue synthetische Daten generiert, während CTGAN und TVAE je nach Konfiguration Schwierigkeiten haben können, neue Datenpunkte zu erzeugen.

Ein möglicher Grund für diesen Unterschied liegt in der zugrunde liegenden Modellarchitektur: PrivBayes nutzt ein probabilistisches Modell, das die Wahrscheinlichkeitsverteilungen der Originaldaten erlernt und darauf basierend neue Datenpunkte erzeugt. Im Gegensatz dazu verlassen sich CTGAN und TVAE auf neuronale Netze, die möglicherweise stärker an die Trainingsdaten gebunden sind und dadurch weniger flexibel bei der Erzeugung neuer, einzigartiger Werte sind.

Zusammenfassend verdeutlicht die grafische Darstellung, dass PrivBayes sich insbesondere durch seine hohe Konsistenz und Stabilität in den verschiedenen Metriken von CTGAN und

TVAE abhebt. Während bei CTGAN und TVAE deutliche Schwankungen in der Modellperformance zu beobachten sind, zeigt PrivBayes über alle Metriken hinweg eine gleichmäßig hohe Leistung. Diese Ergebnisse unterstreichen die bereits zuvor getroffenen Schlussfolgerungen und stützen die Entscheidung, PrivBayes als besonders leistungsfähige Modellfamilie hervorzuheben.

4.6 Zusammenfassung der Ergebnisse

In dieser Arbeit wurden verschiedene Modellkonfigurationen synthetischer Datengenerierung untersucht. Die Analyse erfolgte anhand spezifischer Metriken, die eine Bewertung der statistischen Struktur, Variabilität und Datenschutzqualität der synthetischen Daten ermöglichten. Zudem wurden leistungsstarke Konfigurationen identifiziert und ein Vergleich zwischen den Modellfamilien PrivBayes, CTGAN und TVAE durchgeführt.

4.6.1 Erhaltung der statistischen Struktur

PrivBayes bewahrt die statistischen Eigenschaften der Originaldaten zuverlässig. Die Metriken *Weighted Statistics*, *Column Pair Trends* und *Boundary Adherence* erzielten durchweg hohe Werte.

- **Weighted Statistics:** Die univariaten Verteilungen wurden mit hoher Genauigkeit repliziert. Die Wahl von $k = 2$ führte zu stabileren Ergebnissen als $k = 3$, während der Einfluss des ϵ-Werts gering blieb.
- **Column Pair Trends:** Die paarweisen Abhängigkeiten zwischen Merkmalen wurden gut reproduziert. $k = 2$ erzielte im Durchschnitt leicht bessere Ergebnisse als $k = 3$. Höhere ϵ-Werte verbesserten die Ergebnisse.
- **Boundary Adherence:** Die Wertebereiche der Originaldaten wurden in allen anwendbaren Fällen exakt eingehalten.

4.6.2 Reproduktion der Variabilität der Originaldaten

PrivBayes erhält die Variabilität der Originaldaten weitgehend, wobei die Ergebnisse je nach Datensatz und Parameterauswahl leicht variieren.

- **Value Coverage:** Die meisten Konfigurationen erreichten hohe Werte über 0.95. $k = 2$ lieferte geringfügig bessere Ergebnisse als $k = 3$. Der Einfluss von ϵ war gering, wobei $\epsilon = 1.0$ die besten Ergebnisse erzielte.
- **New Row Synthesis:** Die Erzeugung neuer, nicht direkt aus den Originaldaten stammender Zeilen war bei kleineren Datensätzen höher. Der Einfluss von k war minimal, während höhere ϵ-Werte zu einer geringeren Anzahl neuer Zeilen führten.

4.6.3 Datenschutz und Schutz sensibler Informationen

Der Schutz sensibler Informationen wurde anhand der *Disclosure Protection* bewertet.

- **Disclosure Protection:** Kleinere ϵ-Werte führten zu höheren Schutzwerten, während größere ϵ-Werte die Streuung erhöhten und die Schutzqualität reduzierten. $k = 2$ erzielte im Durchschnitt höhere Werte als $k = 3$, allerdings mit größerer Varianz.

4.6.4 Identifikation leistungsstarker Modellkonfigurationen

Zwei PrivBayes-Konfigurationen erfüllten die Anforderung, in allen Metriken Werte über 0.9 zu erreichen.

- Beide verwendeten $k = 2$ und 10 *bins*, während der ϵ-Wert zwischen 0.1 und 0.5 variierte.
- Die Konfiguration mit $\epsilon = 0.5$ erzielte bessere Ergebnisse in der Reproduktion statistischer Strukturen.
- Die Konfiguration mit $\epsilon = 0.1$ erreichte höhere Werte in der *Disclosure Protection*.

4.6.5 Vergleich der Modellfamilien

Die Modellfamilien PrivBayes, CTGAN und TVAE wurden hinsichtlich ihrer Metrikwerte verglichen.

- **PrivBayes** erzielte konsistent hohe Werte über alle Metriken hinweg. Besonders bei *Column Pair Trends, New Row Synthesis, Value Coverage und Weighted Statistics* lagen die Werte über 0.95.
- **CTGAN und TVAE** zeigten stärkere Schwankungen in den Metriken. Insbesondere bei *Disclosure Protection* fielen einige Werte unter 0.8.
- **New Row Synthesis** variierte bei CTGAN und TVAE stärker, während PrivBayes durchgehend hohe Werte erreichte.

5 Diskussion

In diesem Kapitel werden die zentralen Erkenntnisse aus den Ergebnissen mit bestehenden wissenschaftlichen Arbeiten verglichen. Ziel ist es, herauszuarbeiten, inwieweit die gewonnenen Erkenntnisse mit bereits bekannten theoretischen und empirischen Ansätzen aus der Literatur übereinstimmen oder ob Abweichungen erkennbar sind. Dabei wird untersucht, ob bestehende Forschungsansätze bestätigt werden oder in einem neuen Kontext betrachtet werden können.

Ein besonderer Fokus liegt auf der Wahl des Parameters k in Bayesschen Netzwerken für die Generierung synthetischer Daten, dem Trade-off zwischen Datenschutz und Datenqualität durch die Wahl des ε-Werts, sowie der Repräsentation seltener Werte in den synthetischer Daten. Die Diskussion dieser Aspekte erfolgt unter Berücksichtigung relevanter wissenschaftlicher Arbeiten, um eine fundierte Einordnung der Ergebnisse zu ermöglichen.

5.1 Wahl des k-Parameters

Ein zentraler Aspekt bei der Konstruktion von Bayesschen Netzwerken ist die Wahl der Anzahl abhängiger Knoten k, also der maximalen Anzahl von Elternknoten, die ein Knoten haben kann. Diese Wahl hat direkten Einfluss auf die Qualität der erzeugten synthetischen Daten.

Eine zu hohe Anzahl von Elternknoten kann zunächst den Eindruck erwecken, dass mehr Abhängigkeiten aus den Originaldaten erfasst werden und das synthetische Modell damit eine realistischere Struktur erhält. Allerdings bringt eine hohe Wahl von k mehrere Herausforderungen mit sich: Erstens steigt die rechnerische Komplexität erheblich, da das Finden eines optimalen Netzwerks mit zunehmendem k als NP-schwer gilt [12]. Dies bedeutet, dass sich der Raum möglicher Netzwerkstrukturen mit jeder Erhöhung des Parameters k erheblich vergrößert. Eine ähnliche Problematik wurde auch von Rubio und Gámez [20] beobachtet, die zeigen, dass höhere Werte von k zu einer erheblichen Verschlechterung der Effizienz führen können.

Zweitens wächst mit steigender Anzahl der Elternknoten die Dimension der zu anonymisierenden marginalen Verteilungen, was wiederum die Sensitivität gegenüber Rauschen in differenziell-privaten Mechanismen erhöht [30].

Zhang et al. [30] zeigen in ihrer Arbeit zu PrivBayes, dass eine zu hohe Wahl von k insbesondere dann problematisch wird, wenn das Privacy-Budget ε klein ist. In diesem Fall führt eine zu große Modellkomplexität dazu, dass synthetische Daten verstärkt zufällige Störungen

enthalten, da hochdimensionale Verteilungen durch das hinzugefügte Rauschen verzerrt werden. Dies hat zur Folge, dass die synthetischen Daten zwar formale Datenschutzanforderungen erfüllen, ihre statistische Aussagekraft jedoch sinkt.

Die Wahl von k erfordert daher eine ausgewogene Abwägung zwischen der Erfassung relevanter Abhängigkeiten und der Vermeidung übermäßiger Modellkomplexität. Während ein zu kleines k möglicherweise nicht genug Abhängigkeiten erfasst und dadurch die synthetischen Daten zu stark vereinfacht, führt ein zu großes k zu einer übermäßigen Verzerrung der erzeugten Wahrscheinlichkeitsverteilungen durch das Datenschutzrauschen.

Die in dieser Arbeit durchgeführten Experimente zeigen zudem, dass eine höhere Anzahl von Elternknoten nicht nur zu einer theoretisch höheren Modellkomplexität führt, sondern auch die Berechnungszeit für die Generierung synthetischer Daten signifikant erhöht. Für $k = 2$ konnte die Berechnung in vergleichsweise kurzer Zeit abgeschlossen werden, während sie sich bei $k = 3$ deutlich verlängerte. Zwar lässt sich auf Basis dieser zwei Messwerte keine verlässliche Aussage über den exakten funktionalen Zusammenhang treffen – insbesondere eine exponentielle Approximation wäre an dieser Stelle spekulativ –, jedoch ist eine klare Tendenz zu einer steigenden Rechenzeit erkennbar. Um das Laufzeitverhalten präziser zu analysieren, wären zusätzliche Experimente mit weiteren k-Werten notwendig. Eine Darstellung der gemessenen Laufzeiten ist im Anhang (siehe A.4) zu finden.

In der Praxis bedeutet dies, dass hohe Werte von k nicht nur die Berechnungsdauer erheblich verlängern, sondern auch die Anwendbarkeit des Verfahrens einschränken können. Insbesondere bei großen Datensätzen kann dies zu einem limitierenden Faktor werden. Daher sollte bei der Wahl von k stets ein Abgleich zwischen Modellkomplexität und praktischer Umsetzbarkeit erfolgen.

Ebenfalls zeigen die Experimente in dieser Arbeit, dass ein moderates k bessere Ergebnisse lieferte. Konkret zeigte sich, dass bei $k = 2$ die Qualität der synthetischen Daten, anhand der gemessenen Metrikwerte im Durchschnitt, höher war als bei $k = 3$. Dies steht im Einklang mit den Erkenntnissen aus der Literatur, die darauf hinweisen, dass eine zu hohe Anzahl an Elternknoten nicht zwangsläufig zu einer besseren Modellierung führt, sondern in bestimmten Fällen sogar eine Verschlechterung der Datenqualität verursachen kann.

Zusammenfassend stimmen die experimentellen Ergebnisse mit der in der Literatur beschriebenen Problematik einer zu hohen Modellkomplexität überein und deuten darauf hin, dass eine moderate Wahl von k vorteilhafter sein kann. Sowohl was die Datenqualität, als auch die Rechenzeit für die eigentliche Synthese angeht.

5.2 Trade-off zwischen Datenschutz und Datenqualität durch die Wahl des ε-Werts

Der Privacy-Utility Trade-off wird in dieser Arbeit anhand mehrerer Metriken deutlich, die zur Bewertung der Qualität synthetischer Daten herangezogen werden. Ein Teil der Qualität wird

durch die Metrik Weighted Statistics gemessen, welche die Übereinstimmung der univariaten Verteilungen zwischen synthetischen und Originaldaten beschreibt. Diese Metrik entspricht der Statistical Score-Metrik, die in der Arbeit von Sarmin et al. [21] verwendet wird. Da jedoch unterschiedliche Qualitätsaspekte berücksichtigt werden können, ergänzen weitere Metriken wie Column Pair Trends und New Row Synthesis die Analyse.

Wie in der Arbeit von Sarmin et al. [21] beschrieben, verbessert sich die Datenqualität mit zunehmendem ε, da weniger Rauschen hinzugefügt wird. Die in dieser Arbeit durchgeführten Experimente zeigen jedoch, dass dieser Zusammenhang nicht für alle Datensätze gleichermaßen gilt. Während bei den Datensätzen *Adult* und *News* eine leichte Verbesserung der Weighted Statistics mit steigendem ε beobachtet werden konnte, ließ sich dieser Trend bei anderen Datensätzen nicht in gleicher Weise bestätigen. Darüber hinaus zeigte sich, dass bestimmte Metriken, wie die Boundary Adherence, weitgehend unabhängig vom gewählten ε-Wert blieben.

Ein weiteres Kriterium zur Bewertung der synthetischen Datenqualität ist die Korrelation zwischen Attributen, welche durch die Metrik Column Pair Trends erfasst wird. Diese misst, inwiefern die Beziehungen zwischen Attributen im synthetischen Datensatz erhalten bleiben. Auch hier zeigte sich bei den Datensätzen *Adult* und *Insurance* eine Verbesserung mit höherem ε, während dieser Trend in den übrigen Datensätzen nicht eindeutig erkennbar war. Dies legt nahe, dass eine stärkere Anonymisierung (niedriges ε) dazu führen kann, dass wesentliche statistische Zusammenhänge zwischen Attributen verloren gehen.

Besonders deutlich zeigte sich der Einfluss von ε in der Metrik New Row Synthesis, die erfasst, wie viele der generierten synthetischen Zeilen nicht direkt aus den Originaldaten stammen. Die Ergebnisse zeigen, dass mit steigendem ε die Anzahl neuer Zeilen abnimmt, insbesondere in den Datensätzen *Adult*, *Insurance* und *Child*. Dies deckt sich mit der theoretischen Annahme, dass synthetische Daten mit hohem ε weniger Zufallsvariationen enthalten und stärker durch bestehende Strukturen in den Originaldaten geprägt sind. Dies ist eine direkte Konsequenz des Privacy-Utility Trade-offs: Ein stärkerer Schutz durch niedriges ε führt dazu, dass mehr zufällige Variationen in die Daten eingefügt werden, wodurch sich der Anteil neuartiger Einträge erhöht.

Zusammenfassend zeigen die Ergebnisse dieser Arbeit, dass der Privacy-Utility Trade-off zwar grundsätzlich existiert, seine Auswirkungen jedoch stark von der jeweiligen Metrik und vom zugrunde liegenden Datensatz abhängen. Während sich bei Weighted Statistics und Column Pair Trends der erwartete positive Einfluss von ε auf die Datenqualität nur in bestimmten Datensätzen nachweisen ließ, zeigte sich in der Metrik New Row Synthesis ein klarer Trend zu weniger neuartigen Zeilen bei höheren Werten von ε. Gleichzeitig traten auch Metriken wie Boundary Adherence auf, bei denen sich kein nennenswerter Einfluss beobachten ließ. Diese Ergebnisse legen nahe, dass die Wahl des Privacy-Budgets nicht nur die allgemeine statistischen Eigenschaften, sondern auch die Musterbildung und Neuartigkeit synthetischer Daten in unterschiedlichem Maß beeinflusst.

Einordnung der Disclosure Protection

Die Metrik *Disclosure Protection* misst, inwieweit ein synthetischer Datensatz davor schützt, dass sensible Attribute aus bereits bekannten Informationen erschlossen werden können. Hierzu wird ein standardisiertes Angriffsszenario simuliert, bei dem ein Angreifer über Vorwissen zu bestimmten Spaltenwerten des realen Datensatzes verfügt und versucht, mithilfe des vollständigen synthetischen Datensatzes zusätzliche Merkmale zu erraten. Als Schätzverfahren kommt das sogenannte *CAP*-Verfahren zum Einsatz. Der erzielte Schutzwert wird anschließend mit einer Baseline verglichen, bei der der Angreifer lediglich auf zufällig generierte Daten zugreift. Ein DisclosureProtection-Wert von beispielsweise 0,86 bedeutet, dass der synthetische Datensatz etwa 86% des Schutzes bietet, den komplett zufällige Daten leisten würden.

Für den Adult-Datensatz wurden in den vorliegenden Ergebnissen durchgehend stabile Werte im Bereich von etwa 0,86 beobachtet, was auf eine insgesamt gute Schutzwirkung im Rahmen des beschriebenen Szenarios schließen lässt. Dabei ist zu beachten, dass sich die Metrik auf ein spezifisches, angenommenes Angriffsmuster stützt und maßgeblich von der Auswahl der bekannten und sensiblen Spalten abhängt. Weitere potenzielle Risiken wie Membership-Inference-Angriffe oder die Rekonstruktion seltener Merkmalskombinationen werden von der Metrik nicht explizit abgedeckt. Auch worst-case-Betrachtungen oder der Einfluss zusätzlicher Hintergrundinformationen bleiben unberücksichtigt.

Insgesamt bietet die Metrik eine hilfreiche Orientierung zur Bewertung der Datenschutzwirkung synthetischer Daten, sollte jedoch im Kontext ihrer Annahmen interpretiert werden. Für eine weiterführende Analyse wären ergänzende Metriken oder theoretisch fundierte Schutzgarantien denkbar. PrivBayes verfolgt bereits einen differenziellen Ansatz, der auf dem Konzept der Differential Privacy basiert. Die DisclosureProtection-Metrik prüft jedoch nicht direkt, inwiefern die theoretischen Garantien dieses Konzepts auch unter dem hier simulierten Angriffsszenario greifen.

5.3 Repräsentation seltener Werte in synthetischen Daten

Ein entscheidender Aspekt bei der Generierung synthetischer Daten mit Bayesschen Netzwerken ist die Frage, inwieweit seltene Werte und Randfälle angemessen reproduziert werden. Besonders bei Anwendungen wie INSIGHT, bei denen synthetische Daten zur Testdatengenerierung genutzt werden, kann eine Unterrepräsentation solcher Randfälle problematisch sein.

Bayessche Netzwerke modellieren Wahrscheinlichkeiten auf Basis beobachteter Häufigkeiten in den Originaldaten. Dadurch werden dominante Muster bevorzugt synthetisiert, während seltene Ereignisse tendenziell vernachlässigt werden. Shyalika et al. [23] betonen, dass probabilistische Modelle wie Bayessche Netzwerke seltene Ausprägungen oft nicht adäquat abbilden, da bedingte Wahrscheinlichkeiten vorwiegend durch häufige Kombinationen bestimmt werden.

Für testgetriebene Anwendungen ist dies besonders relevant, da das Fehlen seltener Kombinationen dazu führen kann, dass bestimmte Testszenarien nicht abgedeckt werden.

Ein weiterer Faktor ist, wie Wahrscheinlichkeiten im Netzwerk gelernt werden. Da seltene Werte in den Originaldaten kaum auftreten, fließen sie beim Strukturlernen nur schwach ein, was ihre spätere Reproduktion erschwert oder zu Verzerrungen führt.

Die in dieser Arbeit gemessenen Metrikwerte zur *Value Coverage* könnten auf dieses Phänomen hinweisen: Die synthetischen Daten erreichen zwar eine hohe Abdeckung der in den Originaldaten vorkommenden Werte, jedoch nie einen Wert von 1,0. Dies lässt vermuten, dass insbesondere seltene Werte möglicherweise nicht vollständig abgebildet werden. Eine fundierte Analyse dieses Zusammenhangs wurde im Rahmen dieser Arbeit jedoch nicht durchgeführt und könnte Gegenstand zukünftiger Untersuchungen sein.

Zusammenfassend zeigt sich, dass Bayessche Netzwerke dazu neigen können, seltene Werte zu unterrepräsentieren. Für Anwendungen, bei denen die realistische Abbildung seltener Ereignisse entscheidend ist, stellt dies eine zentrale Herausforderung dar. Eine mögliche Lösung könnte darin bestehen, gezielte Mechanismen zur Verstärkung seltener Kombinationen zu integrieren oder die Netzwerkstruktur entsprechend anzupassen.

6 Zusammenfassung und Konklusion

Die vorliegende Arbeit beschäftigte sich mit der Evaluierung des PrivBayes-Algorithmus zur Erzeugung synthetischer Daten unter Wahrung der Privatsphäre. Ziel war es, die Qualität synthetischer Datensätze, die mithilfe von PrivBayes erzeugt wurden, systematisch zu bewerten und ihre potenzielle Eignung für realitätsnahe Testszenarien abzuleiten. Im Fokus standen dabei neben der Reproduktion statistischer Strukturen auch die Variabilität der Daten und der Schutz sensibler Informationen. Darüber hinaus wurde untersucht, ob sich Konfigurationen identifizieren lassen, die über alle verwendeten Datensätze hinweg konsistente und gute Ergebnisse erzielen.

Im Anschluss wurde die leistungsstärkste Konfiguration von PrivBayes als Referenzpunkt herangezogen, um CTGAN und TVAE auf derselben Datengrundlage und mit identischen Bewertungsmethoden zu analysieren. Dadurch konnte die Leistungsfähigkeit von PrivBayes im Vergleich zu modernen generativen Modellfamilien eingeordnet werden.

Zur abschließenden Einordnung der Ergebnisse werden in diesem Kapitel die fünf Forschungsfragen der Arbeit systematisch beantwortet. Dabei werden zentrale Befunde aus den vorangegangenen Kapiteln gebündelt. In einem abschließenden Abschnitt folgen konkrete Empfehlungen für künftige Arbeiten, die an die hier gewonnenen Erkenntnisse anknüpfen und offene Aspekte weiter vertiefen können.

Forschungsfrage 1: Wie gut erhält PrivBayes die statistischen Eigenschaften der Originaldaten?

Die statistische Struktur der generierten Daten wurde mithilfe der Metriken **Weighted Statistics**, **Boundary Adherence** und **Column Pair Trends** bewertet. Die Ergebnisse zeigen, dass PrivBayes univariate Verteilungen und beobachtete Wertebereiche zuverlässig reproduziert. Die Scores für **Weighted Statistics** lagen über alle Datensätze hinweg konstant auf hohem Niveau (meist über 0,9) und zeigten sich robust gegenüber Veränderungen des ε-Werts. **Boundary Adherence** wies bei allen Datensätzen durchgehend den Maximalwert auf, was auf eine realitätsnahe Werteerzeugung innerhalb der beobachteten Grenzen hinweist.

Die Metrik **Column Pair Trends** zeigte hingegen größere Streuung. Zwar wurden auch hier teils hohe Werte erreicht, allerdings variierte die Reproduzierbarkeit paarweiser Attributbeziehungen stärker zwischen den Datensätzen und Konfigurationen. Ein Zusammenhang zwischen höheren ε-Werten und besseren Scores ließ sich beobachten. Auch der Parameter k, der die

maximale Anzahl an Elternknoten im Bayesschen Netzwerk bestimmt, hatte einen erkennbaren Einfluss: Konfigurationen mit $k = 2$ führten tendenziell zu stabileren und höheren Scores als solche mit $k = 3$.

Diese Ergebnisse deuten darauf hin, dass PrivBayes in der Lage ist, zentrale statistische Eigenschaften der Originaldaten zu bewahren. Die im Diskussionsteil herausgearbeiteten Effekte der Modellkomplexität und der Privatsphärenstärke werden durch die empirischen Befunde gestützt.

Forschungsfrage 2: Wie gut reproduziert PrivBayes die Variabilität der Originaldaten?

Zur Beantwortung dieser Frage wurden die Metriken **Value Coverage** und **New Row Synthesis** betrachtet. Die Ergebnisse zeigen, dass PrivBayes eine hohe Vielfalt erzeugt, indem es neue Kombinationen bestehender Attributausprägungen generiert. Die **Value Coverage** erreichte in allen getesteten Konfigurationen Werte zwischen 0,91 und 1,0. Damit wird bestätigt, dass PrivBayes die beobachteten Kategorien und numerischen Wertebereiche des Originals nahezu vollständig abdeckt.

Dabei war für den Adult-Datensatz die *Category Coverage* durchweg hoch, da PrivBayes ausschließlich auf den im Training beobachteten Kategorien synthetisierte. Bei der *Range Coverage* traten in einzelnen Konfigurationen geringfügige Lücken auf, die darauf hindeuten, dass nicht alle Zwischenwerte eines numerischen Intervalls durchgängig reproduziert wurden. Da die Metrik bereits kleine Lücken negativ bewertet, ist eine vollständige Abdeckung kontinuierlicher Skalen methodisch anspruchsvoller als bei kategorialen Merkmalen.

Die Metrik **New Row Synthesis** zeigte ebenfalls hohe Werte. Dies belegt, dass PrivBayes synthetische Zeilen erzeugt, die nicht vollständig mit einzelnen Zeilen aus dem Original übereinstimmen. Besonders bei niedrigen Werten von ε wurden tendenziell neuartigere Kombinationen erzeugt. Bei höheren ε-Werten nahm die Vielfalt hingegen leicht ab, was den in der Diskussion beschriebenen Zielkonflikt zwischen Datenschutz und generativer Vielfalt empirisch stützt.

Insgesamt lässt sich festhalten, dass PrivBayes keine neuen Kategorien oder Werte erfindet und durch die Rekombination bekannter Ausprägungen eine hohe strukturelle Variabilität erreicht.

Forschungsfrage 3: Inwieweit gewährleistet PrivBayes den Schutz sensibler Informationen?

Zur Bewertung der Schutzwirkung wurde die Metrik Disclosure Protection verwendet, die ein realistisches Angriffsszenario simuliert. Dabei wird geprüft, ob ein Angreifer auf Basis bekannter Attribute wie Alter und Geschlecht sensible Informationen wie politische Einstellungen vorhersagen kann. Der berechnete Risiko-Score wird mit einer Zufallsbaseline verglichen.

Die Auswertung ergab, dass Konfigurationen mit maximal zwei Elternknoten (k=2) höhere Schutzwerte aufwiesen als solche mit drei Elternknoten (k=3). Auch beim Privatsphärenparameter ε zeigte sich eine Tendenz: Besonders niedrige Werte erzielten die höchsten Schutzwerte. Mit zunehmendem ε sank der Score zunächst deutlich, bei $\varepsilon = 10{,}0$ stieg er jedoch leicht an. Insgesamt lässt sich festhalten, dass PrivBayes unter datenschutzfreundlichen Einstellungen (kleines ε) in der Lage ist, Rückschlüsse auf sensible Informationen effektiv zu erschweren.

Forschungsfrage 4: Gibt es eine Konfiguration, die über alle Datensätze hinweg gute Ergebnisse erzielt?

Zur Beantwortung dieser Frage wurden die Ergebnisse aller fünf Metriken über sämtliche Datensätze hinweg aggregiert und anschließend auf Ebene der Konfigurationen zusammengeführt. Ziel war es, jene Einstellungen zu identifizieren, die unabhängig vom Datensatz eine konstant hohe Qualität synthetischer Daten erzielen.

Auf Basis der aggregierten Ergebnisse schnitt die Konfiguration mit $\varepsilon = 1{,}0$ und $k = 2$ (maximal zwei Elternknoten im Bayesschen Netzwerk) am besten ab. Sie erzielte im Mittel die höchsten Scores über alle betrachteten Qualitätsdimensionen hinweg.

Gleichzeitig zeigte die detaillierte Auswertung einzelner Datensätze, dass auch bei dieser Konfiguration in Einzelfällen Einschränkungen auftraten, insbesondere bei der *Disclosure Protection*, die in manchen Fällen unterhalb von 0,6 lag. Diese Beobachtung unterstreicht, dass aggregierte Bewertungen zwar wichtige Hinweise auf leistungsstarke Grundeinstellungen liefern können, bei der konkreten Anwendung jedoch stets eine individuelle Prüfung der Qualität ratsam ist.

Forschungsfrage 5: Wie unterscheiden sich die Ergebnisse von PrivBayes, CTGAN und TVAE

Zur Beantwortung dieser Frage wurden neben den selbst erzeugten Datensätzen auch solche verwendet, die mithilfe von CTGAN und TVAE generiert und im Rahmen des Projekts INSIGHT von der begleitenden Forschungsgruppe bereitgestellt wurden. Ziel war es, die Leistungsfähigkeit von PrivBayes im Vergleich zu modernen, auf neuronalen Netzen basierenden Generierungsverfahren einzuordnen. Die Bewertung erfolgte anhand derselben Metriken wie im vorherigen Vergleich.

In der Gesamtschau erzielte PrivBayes in allen betrachteten Metriken die höchsten Werte. Die Unterschiede zu CTGAN und TVAE fielen dabei jedoch vergleichsweise gering aus und lagen in vielen Fällen nur wenige Punkte auseinander. Dies deutet darauf hin, dass auch generative Modelle ohne explizite Datenschutzvorgaben in der Lage sind, qualitativ hochwertige synthetische Daten zu erzeugen.

Gleichzeitig ist zu berücksichtigen, dass CTGAN und TVAE nicht darauf ausgelegt sind, differenzielle Privatsphäre zu gewährleisten. Vor diesem Hintergrund ist der Vorsprung von

PrivBayes besonders bemerkenswert, da es unter strikten Datenschutzvorgaben dennoch eine sehr hohe Datenqualität erzielen konnte.

Konklusion

Die Ergebnisse dieser Arbeit zeigen, dass sich mit PrivBayes qualitativ hochwertige synthetische Datensätze erzeugen lassen, die zentrale Eigenschaften der Originaldaten bewahren und zugleich ein hohes Datenschutzniveau erreichen können. Die Wahl der Parameter hatte dabei einen spürbaren Einfluss auf die Ergebnisqualität. Besonders Konfigurationen mit $k = 2$ (maximal zwei Elternknoten) und $\varepsilon = 1{,}0$ erwiesen sich über alle Datensätze hinweg als leistungsstark. Gleichzeitig wurde deutlich, dass auch die Datenstruktur selbst einen maßgeblichen Einfluss hat. Entsprechend empfiehlt es sich, diese Konfiguration als fundierten Ausgangspunkt zu wählen, die finale Parametrierung jedoch stets eigenständig zu evaluieren und bei Bedarf mit alternativen Einstellungen abzugleichen.

Die eingesetzten Metriken bieten eine fundierte Grundlage zur Bewertung synthetischer Daten. Sie erfassen insbesondere statistische Ähnlichkeit, Variabilität und Datenschutzaspekte. Dennoch decken sie nicht alle relevanten Qualitätsdimensionen ab. So lassen sich beispielsweise Aussagen zur Eignung für machine learning (ML) aus diesen Metriken nur bedingt ableiten. Es ist denkbar, dass hohe Scores bei der statistischen Ähnlichkeit nicht zwangsläufig mit einer guten Performance von ML-Modellen einhergehen. Eine starke Schutzwahl (niedriges ε) könnte diese Performance noch weiter abschwächen. Weitere Untersuchungen sollten deshalb prüfen, inwiefern Metriken wie Weighted Statistics oder New Row Synthesis mit ML-Metriken wie Accuracy oder F1 korrelieren oder sich gegenseitig beeinflussen.

Darüber hinaus sollte die Übertragbarkeit der Ergebnisse auf produktive Szenarien weiter untersucht werden. In dieser Arbeit wurden ausschließlich Demo-Datensätze verwendet. Diese erlauben zwar realitätsnahe Einschätzungen, bilden jedoch nicht alle Herausforderungen produktiver Daten ab, vor allem im Hinblick auf Heterogenität, Datenqualität oder regulatorische Anforderungen. Künftige Arbeiten sollten daher prüfen, ob sich die beobachteten Effekte unter realweltlichen Bedingungen bestätigen lassen.

Nicht zuletzt erscheint es sinnvoll, in zukünftigen Vergleichen auch weitere Modellansätze systematisch einzubeziehen. PrivBayes bietet durch sein bayessches Netzwerk in Kombination mit differenzieller Privatsphäre eine robuste und erklärbare Lösung, ist jedoch hinsichtlich der unterstützten Datentypen eingeschränkt. DP-GAN [27] und PATE-GAN [29] stellen zwei alternative Modellansätze dar, die differenzielle Privatsphäre durch den Einsatz generativer neuronaler Netze umsetzen. Während DP-GAN besonders für strukturierte, hochdimensionale Daten geeignet ist und den Datenschutz durch direktes Rauschen im Trainingsprozess gewährleistet, setzt PATE-GAN auf mehrere getrennt trainierte Modelle, deren gemeinsame Entscheidungen durch gezieltes Rauschen geschützt werden, um die Privatsphäre zu wahren. Beide Modelle bieten Potenzial für den Einsatz in komplexeren Anwendungsszenarien und sollten in zukünftigen Arbeiten hinsichtlich Datenqualität und Privacy-Wirkung untersucht werden.

Literaturverzeichnis

[1] Beduschi, Ana: Synthetic data protection: Towards a paradigm change in data regulation? In: *Big Data & Society* 11 (2024), Nr. 1, S. 20539517241231277. – URL https://doi.org/10.1177/20539517241231277

[2] Binder, John ; Koller, Daphne ; Russell, Stuart ; Kanazawa, Keiji: Adaptive probabilistic networks with hidden variables. In: *Machine Learning* 29 (1997), Nr. 2, S. 213–244

[3] Blackard, Jock A.: *Comparison of neural networks and discriminant analysis in predicting forest cover types.* Colorado State University, 1998

[4] Bowen, Claire M. ; Snoke, Joshua: Comparative Study of Differentially Private Synthetic Data Algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge. In: *Journal of Privacy and Confidentiality* 11 (2021), Feb., Nr. 1. – URL https://journalprivacyconfidentiality.org/index.php/jpc/article/view/748

[5] DataCebo, Inc. (Veranst.): *Synthetic Data Metrics.* 12 2024. – URL https://docs.sdv.dev/sdmetrics/. – Version 0.18.0

[6] Du, Yuntao ; Li, Ninghui: *Systematic Assessment of Tabular Data Synthesis.* 2025. – URL https://openreview.net/forum?id=3ANoEa7roV

[7] Dwork, Cynthia ; Roth, Aaron: The Algorithmic Foundations of Differential Privacy. In: *Foundations and Trends® in Theoretical Computer Science* 9 (2014), Nr. 3–4, S. 211–407. – URL http://dx.doi.org/10.1561/0400000042. – ISSN 1551-305X

[8] Fawzy, Ahmed ; Tahir, Amjed ; Galster, Matthias ; Liang, Peng: Exploring data management challenges and solutions in agile software development: a literature review and practitioner survey. In: *Empirical Software Engineering* 30 (2025), Mar, Nr. 3, S. 77. – URL https://doi.org/10.1007/s10664-025-10630-4. – ISSN 1573-7616

[9] Fernandes, Kelwin ; Vinagre, Pedro ; Cortez, Paulo: A proactive intelligent decision support system for predicting the popularity of online news. In: *Portuguese conference on artificial intelligence* Springer (Veranst.), 2015, S. 535–546

[10] Goyal, Mandeep ; Mahmoud, Qusay H.: A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. In: *Electronics* 13 (2024), Nr. 17. – URL https://www.mdpi.com/2079-9292/13/17/3509. – ISSN 2079-9292

[11] Harbron, James u. a.: Synthetic data use: exploring use cases to optimise data utility. In: *Discover Artificial Intelligence* 1 (2021), Nr. 16. – URL https://link.springer.com/article/10.1007/s44163-021-00016-y

[12] Heckerman, David ; Geiger, Dan ; Chickering, David M.: Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. In: *Machine Learning* 20 (1995), Sep, Nr. 3, S. 197–243. – URL https://doi.org/10.1023/A:1022623210503. – ISSN 1573-0565

[13] Karst, Fabian S. ; Li, Mahei M. ; Leimeister, Jan M.: SynDEc: A Synthetic Data Ecosystem. In: *Electronic Markets* 35 (2025), Jan, Nr. 1, S. 7. – URL https://doi.org/10.1007/s12525-024-00746-8. – ISSN 1422-8890

[14] Kohavi, Ron u. a.: Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In: *Kdd* Bd. 96, 1996, S. 202–207

[15] Kokosi, Theodora ; Harron, Katie: Synthetic data in medical research. In: *BMJ Medicine* 1 (2022), 09. – URL https://doi.org/10.1136/bmjmed-2022-000167

[16] Neekhra, Bhavesh ; Kapoor, Kshitij ; Gupta, Debayan: *Synthpop++: A Hybrid Framework for Generating A Country-scale Synthetic Population*. 2024. – URL https://arxiv.org/abs/2304.12284

[17] Qian, Zhaozhi ; Callender, Thomas ; Cebere, Bogdan ; Janes, Sam ; Navani, Neal ; Schaar, Mihaela: Synthetic data for privacy-preserving clinical risk prediction. In: *Scientific Reports* 14 (2024), 10. – URL https://doi.org/10.1038/s41598-024-72894-y

[18] Qian, Zhaozhi ; Cebere, Bogdan-Constantin ; Schaar, Mihaela van der: *Synthcity: facilitating innovative use cases of synthetic data in different data modalities*. 2023. – URL https://arxiv.org/abs/2301.07573

[19] Rocher, Luc ; Hendrickx, Julien M. ; Montjoye, Yves-Alexandre de: Estimating the success of re-identifications in incomplete datasets using generative models. In: *Nature Communications* 10 (2019), Jul, Nr. 1, S. 3069. – URL https://doi.org/10.1038/s41467-019-10933-3. – ISSN 2041-1723

[20] Rubio, Arcadio ; Gámez, José A.: Flexible learning of k-dependence Bayesian network classifiers. In: *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*. New York, NY, USA : Association for Computing Machinery, 2011 (GECCO '11), S. 1219–1226. – URL https://doi.org/10.1145/2001576.2001741. – ISBN 9781450305570

[21] Sarmin, Fatima J. ; Sarkar, Atiquer R. ; Wang, Yang ; Mohammed, Noman: *Synthetic Data: Revisiting the Privacy-Utility Trade-off*. 2025. – URL https://arxiv.org/abs/2407.07926

[22] Sella, Nadir ; Guinot, Florent ; Lagrange, Nikita ; Albou, Laurent-Philippe ; Desponds, Jonathan ; Isambert, Hervé: Preserving information while respecting privacy through an information theoretic framework for synthetic health data generation. In: *npj Digital Medicine* 8 (2025), 01. – URL https://doi.org/10.1038/s41746-025-01431-6

[23] Shyalika, Chathurangi ; Wickramarachchi, Ruwan ; Sheth, Amit P.: A Comprehensive Survey on Rare Event Prediction. In: *ACM Comput. Surv.* 57 (2024), November, Nr. 3. – URL https://doi.org/10.1145/3699955. – ISSN 0360-0300

[24] Specht, Felix ; Otto, Jens ; Ratz, Daniel: Generation of Synthetic Data to Improve Security Monitoring for Cyber-Physical Production Systems. In: *2023 IEEE 21st International Conference on Industrial Informatics (INDIN)*, URL https://ieeexplore.ieee.org/document/10218171, 2023, S. 1–7

[25] Spiegelhalter, David J.: Learning in probabilistic expert systems. In: *Bayesian statistics* 4 (1992), S. 447–465

[26] Tan, Chao ; Behjati, Razieh ; Arisholm, Erik: A Model-Based Approach to Generate Dynamic Synthetic Test Data: A Conceptual Model. In: *2019 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, URL https://ieeexplore.ieee.org/document/8728950, 2019, S. 11–14

[27] Xie, Liyang ; Lin, Kaixiang ; Wang, Shu ; Wang, Fei ; Zhou, Jiayu: *Differentially Private Generative Adversarial Network*. 2018. – URL https://arxiv.org/abs/1802.06739

[28] Xu, Lei ; Skoularidou, Maria ; Cuesta-Infante, Alfredo ; Veeramachaneni, Kalyan: Modeling Tabular data using Conditional GAN. In: Wallach, H. (Hrsg.) ; Larochelle, H. (Hrsg.) ; Beygelzimer, A. (Hrsg.) ; Alché-Buc, F. d'(Hrsg.) ; Fox, E. (Hrsg.) ; Garnett, R. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 32, Curran Associates, Inc., 2019. – URL https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf

[29] Yoon, Jinsung ; Jordon, James ; Schaar, Mihaela van der: PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. In: *International Conference on Learning Representations*, URL https://openreview.net/forum?id=S1zk9iRqF7, 2019

[30] Zhang, Jun ; Cormode, Graham ; Procopiuc, Cecilia M. ; Srivastava, Divesh ; Xiao, Xiaokui: PrivBayes: Private Data Release via Bayesian Networks. In: *ACM Trans. Database Syst.* 42 (2017), Oktober, Nr. 4. – URL https://doi.org/10.1145/3134428. – ISSN 0362-5915

A Anhang

Exemplarischer Python-Code zur Generierung synthetischer Daten

```python
import pandas as pd
from synthcity.plugins import Plugins

# Originaldaten laden
data = pd.read_csv("Demodaten/news.csv")

# Stichprobe von 500 Zeilen ziehen
sampled_data = data.sample(n=500, random_state=42)

# Parameter für PrivBayes
epsilon_values = [0.1, 0.5, 1, 10]
bins_values = [10, 20]
k_values = [2, 3]

# Generierung durchführen
for epsilon in epsilon_values:
    for bins in bins_values:
        for k in k_values:
            # Modell initialisieren
            model = Plugins().get(
                "privbayes",
                epsilon=epsilon,
                n_bins=bins,
                K=k
            )

            # Training und Generierung
            model.fit(sampled_data)
            synthetic_data = model.generate(count=len(sampled_data)).dataframe()

            # Ergebnisse speichern
            file_name = f"news/synthetic_data_news500_e{epsilon}_b{bins}_k{k}.csv"
            synthetic_data.to_csv(file_name, index=False)
            print(f"Datei gespeichert: {file_name}")
```

Abbildung A.1: Python-Code zur Anwendung von PrivBayes

Plots zur Beantwortung der Forschungsfragen und der Diskussion

Plots, die zur visuellen Untermauerung der Forschungsfragen beitragen und der Diskussion.

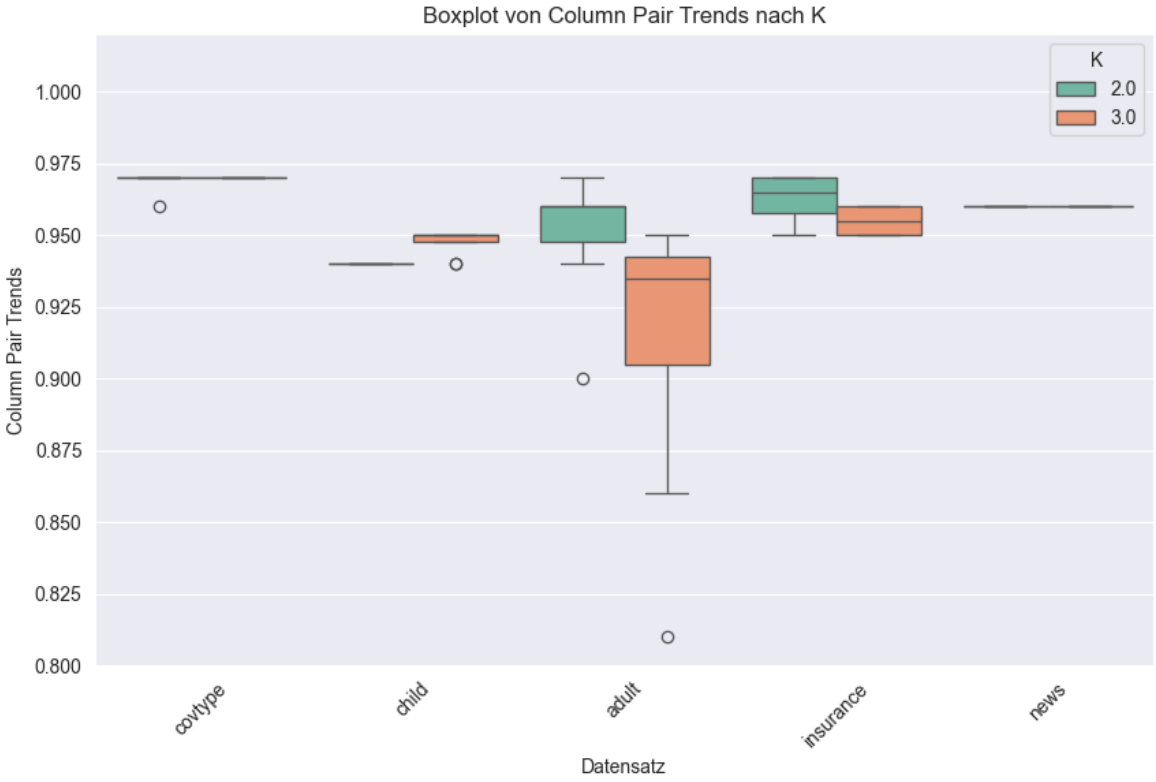


Abbildung A.2: Column Pair Trends für alle Datensätze (K = 2 vs. K = 3)

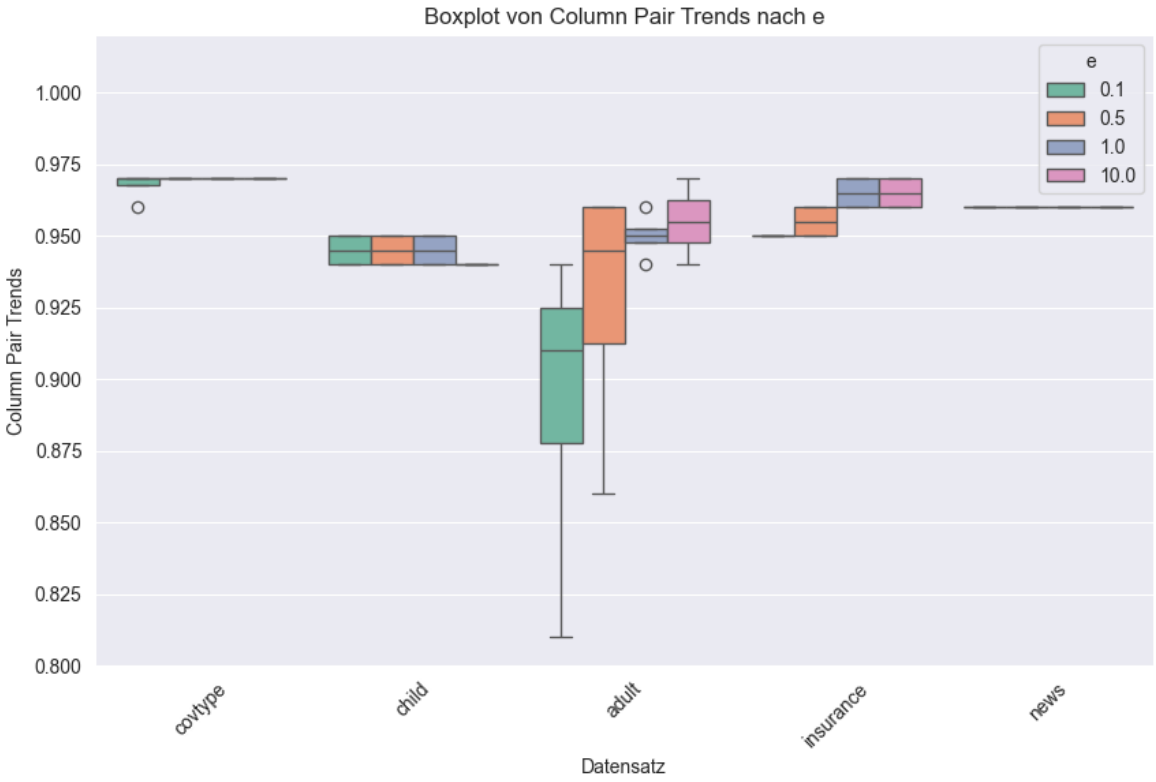


Abbildung A.3: Column Pair Trends für alle Datensätze über alle unterschiedlichen ϵ

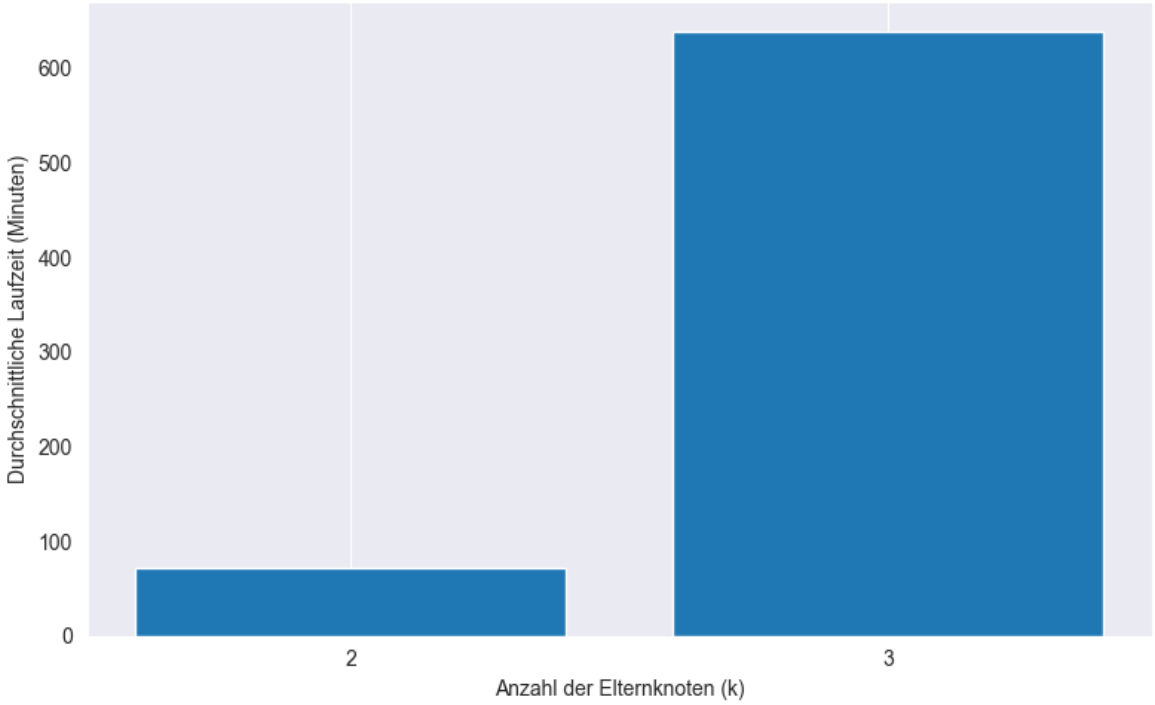


Abbildung A.4: Durchschnittliche Laufzeit pro k-Wert für den Covtype-Datensatz

Erklärung zur selbständigen Bearbeitung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

Ort	Datum	Unterschrift im Original