technische universität dortmund

IT in Produktion und Logistik

Experimenteller Vergleich von Verfahren der Datentransformation und Datenreduktion im Kontext von Data Mining

Masterarbeit zur Erlangung des Grades M. Sc.

Vorgelegt von:	Juliana Xenia Kropf
Matrikelnummer:	166732
Studiengang:	Maschinenbau

Ausgabedatum:	06.01.2025
Abgabedatum:	23.06.2025

Erstprüfer:	Dr.-Ing. Anne Antonia Scheidler
Zweitprüfer:	Florian Hochkamp, M. Sc.

Technische Universität Dortmund

Fakultät Maschinenbau

Fachgebiet IT in Produktion und Logistik

Inhaltsverzeichnis

1 Einleitung

Durch Informations- und Kommunikationstechnologien entstehen in nahezu allen Bereichen der Industrie enorme Datenmengen, die als Rohstoff für wertvolle Erkenntnisse dienen (Bauernhansl 2023). Im Zeitalter der Industrie 4.0 sind Unternehmen und Forschung zunehmend auf datengetriebene Entscheidungen angewiesen (Bauernhansl 2023). Die Fortschritte der letzten fünf Jahre zeigen hierin eine rapide Entwicklung in Bereichen der Parallelisierungstechniken und Cloud Computing (Yesilyurt et al. 2023), wobei zudem bessere Hardware zur Datenverarbeitung zur Verfügung steht (Challa et al. 2024). Die global erzeugt, gespeicherte und genutzte Datenmenge wird Prognosen zufolge von 149 Zettabytes im Jahr 2024 bis 2028 auf 394 Zettabytes ansteigen (Statista 2024). Eine nahezu Verdreifachung innerhalb von nur vier Jahren, die die kritische Notwendigkeit effizienten Datenmanagements in allen Bereichen des digitalen Wandels unterstreicht. Einfache Verfahren zur Auswertung dieser Datenmengen genügen inzwischen nicht mehr. Insbesondere ist die Datenqualität häufig unzureichend, um valide Erkenntnisse zu generieren (Han 2012). Genau hier setzt die Datenvorverarbeitung an, die nicht nur die Quantität, sondern auch die Qualität der Daten optimiert, um deren Analysepotenzial voll auszuschöpfen. Um die wachsende Datenmenge effizient zu verarbeiten, bieten Verfahren der Datenreduktion eine Lösung. Jede Reduktion ist jedoch mit dem Risiko verbunden, unwiederbringlich wertvolle Daten und Informationen zu verlieren. Weiter noch können Rohdaten unvollständig oder unstrukturiert vorliegen und in inkompatiblen Intervallen skaliert sein (García et al. 2015). Ein Problem, dass ohne eine systematische Aufarbeitung zu unbrauchbaren Ergebnissen führt. Die eigentliche Herausforderung hierbei liegt in ihrer Transformation zu handlungsrelevantem Wissen, welches das Potenzial besitzt Entscheidungsprozesse zu optimieren (Thearling 2010; Linoff und Berry 2011). Durch Verfahren der Datentransformation werden Datensätze in eine geeignete Form für einen entsprechenden Data-Mining-Algorithmus überführt (Linoff und Berry 2011). Ohne ausreichendes Domänenverständnis kann es dabei jedoch zu unbeabsichtigten Verzerrungen kommen. Werden ohne Hintergrundwissen zu den Datensätzen Entscheidungen bei der Verfahrensauswahl in der Datenvorverarbeitung getroffen, so beeinträchtigt dies direkt die Güte der Analyseergebnisse (Thearling 2010). Damit diese Analysen durchgeführt werden können, müssen sie von geeigneten, möglichst automatisierten Verfahren unterstützt werden (Linoff und Berry 2011). Genau hier setzt die Wissensentdeckung in Datenbanken an, die mithilfe systematischer Prozesse und einen entsprechenden Data-Mining-Algorithmus verborgene Muster und Zusammenhänge in Daten identifiziert (Han 2012; Linoff und Berry 2011). Als integraler Bestandteil dieses Prozesses, adressiert die Datenvorverarbeitung sowohl die Transformation als auch die Reduktion von Daten, um eine Grundlage für valide und belastbare Ergebnisse zu schaffen.

Für den Prozess der Wissensentdeckung in Datenbanken existieren etablierte Vorgehensmodelle, wie das Vorgehensmodell nach Fayyad et al. (1996), die strukturierte Phasen von der Aufbereitung der Rohdaten zu Informationen bis hin zur Extraktion von Wissen definieren. In der Datenvorverarbeitung sind die Schritte der Datentransformation und Datenreduktion entscheidend, um Datenqualität sicherzustellen, Dimensionalität zu reduzieren und die Skalierbarkeit nachgelagerter Data-Mining-Schritte zu gewährleisten (Linoff und Berry 2011).

Das Hauptziel dieser Arbeit ist es, basierend auf etablierter Literatur zu Verfahren der Datentransformation und Datenreduktion einen Experimentierplan zu entwickeln, der einen systematischen Vergleich nach einem festgelegten Bewertungsschema ermöglicht. Um dieses Ziel zu erreichen, werden drei Teilziele verfolgt. Das erste Teilziel befasst sich damit Evaluationskriterien zu definieren, um die Performanz der Datenvorverarbeitungsverfahren bewertbar zu machen. Im nächsten Teilziel wird ein Experimentierplan entwickelt, der einen strukturierten Rahmen für die vergleichende Analyse der Verfahren schafft. Schließlich erfolgt im letzten Teilziel die exemplarische Verprobung der Erkenntnisse anhand eines Fallbeispiels, um deren praktische Anwendbarkeit zu validieren.

Zur erreichen der Zielsetzung wird die folgende methodische Vorgehensweise angewandt. In Kapitel 2 wird zunächst das konzeptuelle Fundament der Arbeit gesetzt. Hierzu werden die zentralen Begriffe Daten, Informationen und Wissen definiert und voneinander abgegrenzt, um eine einheitliche Terminologie zu etablieren. Anschließend werden Datenbanken erläutert und in diesem Zusammenhang Datenqualität definiert. Darauf aufbauend werden etablierte Vorgehensmodelle der Wissensentdeckung vorgestellt, um den systematischen Rahmen für nachfolgende Analysen zu schaffen. Als Nächstes werden die Grundlagen des Data Mining aufgearbeitet und dessen Rolle im Kontext der Wissensentdeckung beleuchtet. Abschließend wird eine breite Einführung in die Datenvorverarbeitung durchgeführt, die zentrale Techniken und deren Bedeutung für die Datenanalyse herausstellt. Kapitel 3 widmet sich der detaillierten Auseinandersetzung mit den Verfahren der Datentransformation und Datenreduktion. Zunächst werden die Grundlagen der Datentransformation erläutert. Im Anschluss werden konkrete Verfahren der Datentransformation systematisch kategorisiert und deren Anwendungsgebiete behandelt. Analog erfolgt eine Einführung in die Grundlagen der Datenreduktion, gefolgt von der Vorstellung einiger Verfahren der Datenreduktion. Aufbauend auf den theoretischen Grundlagen wird in Kapitel 4 ein Bewertungsschema entwickelt. Anschließend wird ein Experimentierplan erstellt, der die Vergleichbarkeit der Verfahren sicherstellt. In Kapitel 5 erfolgt die praktische Validierung der gewonnenen Erkenntnisse. Ein Fallbeispiel, das repräsentative Daten und Anforderungen aus einer konkreten Domäne nutzt, wird eingeführt. Schließlich werden die Ergebnisse kritisch diskutiert und Rückschlüsse für die Praxis gezogen. Abschließend fast Kapitel 6 die Ergebnisse der Arbeit zusammen. Zudem werden offene Aspekte thematisiert und potenzielle Ansätze für weiterführende Forschungsfragen erläutert.

2 Grundlagen der Wissensentdeckung in Datenbanken

In diesem Kapitel werden die Grundlagen der Wissensentdeckung erläutert. Zunächst wird der Zusammenhang zwischen den Begriffen Daten, Informationen und Wissen dargestellt und eine begriffliche Grundlage hinsichtlich ihrer Definitionen erstellt. Anschließend werden gängige Vorgehensmodelle der Wissensentdeckung vorgestellt und deren Phasen von der Datenselektion bis zur Wissensinterpretation definiert sowie zentrale Merkmale beschrieben. Ein weiterer Fokus liegt auf den Grundlagen der Datenvorverarbeitung, die eine essenzielle Rolle innerhalb des Prozesses der Wissensentdeckung spielt. Abschließend werden Anforderungen aus dem Data Mining beleuchtet, die maßgeblich die Auswahl von Vorverarbeitungsverfahren beeinflussen, die in den folgenden Kapiteln der Arbeit adressiert werden.

2.1 Daten, Informationen und Wissen

Während in der Literatur die Übermenge der Daten als Datenreichtum bei gleichzeitiger Informationsarmut charakterisiert wird (Lewandowski 1999), beschreibt Naisbitt (1982) diese Diskrepanz als ein Ertrinken in Information bei Hungersnot nach Wissen. Die scheinbar widersprüchlichen Metaphern verdeutlichen die grundsätzlich inkonsistenten Begriffsdefinitionen. Eine homogene Definition der Begriffe Daten, Informationen und Wissen ist in der Literatur selbst innerhalb einzelner Fachbereiche nicht einheitlich etabliert (Frederik Möller et al. 2017). Die wachsende Datenmenge, wie in Kapitel 1 beschrieben, erfordert eine zielgerichtete Nutzung und sinnvolle Strukturierung, um ihr Potenzial für neues Wissen voll auszuschöpfen. Begriffliche Unschärfe behindert an dieser Stelle die Vergleichbarkeit von Methoden. Eine präzise, anwendungsfallbezogene Abgrenzung dieser Begriffe ist somit erstrebenswert, um die Brücke zwischen theoretischen Modellen und praktischer Umsetzung zu schlagen.

Die in Abbildung 1 dargestellte Wissenstreppe nach North (2021) findet in wissenschaftlichen Publikationen regelmäßig Anwendung in den Beschreibungen von Daten, Informationen und Wissen.

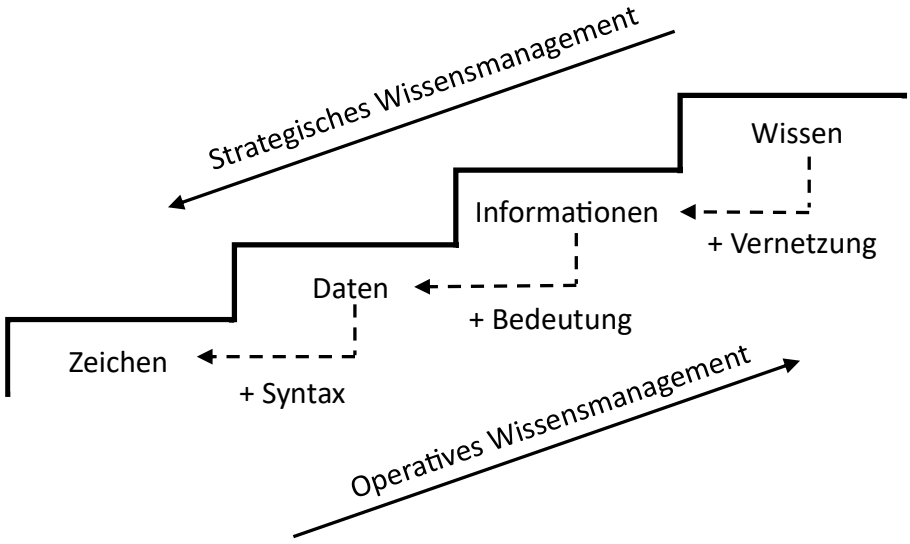

Abbildung 1: North'sche Wissenstreppe (eigene Darstellung in Anlehnung an North (2021))

In Abbildung 1 werden die Daten als Zeichen oder Symbole definiert, die durch Syntaxregeln strukturiert sind (Bodendorf 2006; North 2021). In der Literatur werden Daten auch als uninterpretierte Symbole beschrieben (Ackoff 1989; Floridi 2005). Information entsteht somit durch die strukturierte Kontextualisierung von Daten, während Wissen durch die Vernetzung von Information mit weiterer Information sowie unter Reflexion und Erfahrungsintegration gebildet wird (North 2021). Nur so ist es möglich handlungsrelevante Erkenntnisse abzuleiten.

In dieser Arbeit werden Daten als Symbole ohne Eigenwert verstanden, die durch Merkmale beschrieben und in persistenter Form gespeichert werden können (Ackoff 1989; Laurini 2017). Zusammengehörige Daten können als Datensätze abgelegt werden (Han 2012).

Im Hinblick auf die Datenanalyse lassen sich Daten hinsichtlich ihrer verschiedenen, zu beobachtenden Merkmale erfassen. Es wird zwischen univariaten, bivariaten und multivariaten Merkmalen unterschieden. Diese Klassifikation reflektiert die Komplexität der Analyse. Während univariate Daten einfache Häufigkeitsverteilungen abbilden, ermöglichen multivariate Analysen die Untersuchung komplexer Interaktionen. Daten selbst sind jedoch zunächst lediglich syntaktische Zeichenfolgen ohne inhärente Bedeutung. Weiter werden Informationen als interpretierbare Daten definiert (Zins 2007; Ackoff 1989; North 2021) und entstehen entsprechend erst durch semantische Interpretation (Katharina Schüller et al. 2019; Röhner und Schütz 2016; Frické 2019). Die Wissenstreppe verdeutlicht, dass Wissen erst durch Handlungsrelevanz und Anwendung einen Mehrwert erzeugen kann. Für diese Arbeit wird Wissen als eine nicht-triviale Erkenntnis definiert, die erst durch den Prozess der Datenanalyse und die Interpretation von Information zugänglich wird und sich durch ihre Neuheit auszeichnet. Dieses Verständnis knüpft an North (2021) an, der Wissen als „zweckdienliche Vernetzung von Informationen durch das Bewusstsein" beschreibt. Wissen ist somit stets anwendungsspezifisch und entfaltet seinen Wert erst in seinem praktischen Einsatz. Hierzu beschreiben

Collins und Smith (2006) Wissen zudem als einen Prozess, der die Erfassung, Analyse und Visualisierung von Daten umfasst, mit dem Ziel, unerwartetes Wissen zu generieren. Kernaufgabe ist es, durch strukturierte Datenvorverarbeitung belastbare Erkenntnisse zu schaffen. Ein Prozess, der folglich die klare Trennung von Daten, Informationen und Wissen voraussetzt. Im Zusammenhang mit der Wissensentdeckung in Datenbanken dienen Muster als Brücke zwischen Rohdaten und Erkenntnissen, deren Bedeutung jedoch erst durch menschliche Expertise vollständig erschlossen wird (Maimon und Rokach 2010). Muster zeichnen sich als wiederkehrende Strukturen in Daten ab, die ohne systematische Analyse unerkannt blieben. Im Rahmen der Wissensentdeckung in Datenbanken identifizieren Data-Mining-Algorithmen diese Muster, die jedoch erst durch menschliche Validierung und kontextuelle Einordnung zu Wissen werden (Steiner 2021; Schelter et al. 2018).

Die systematische Erkennung von Mustern setzt nicht nur analytische Methoden voraus, sondern auch die Zugänglichkeit auf gespeicherte Daten. Diese Aufgabe übernehmen Datenbanken, die im folgenden Kapitel 2.2 behandelt werden.

2.2 Datenbanken und Datenqualität

Daten werden in persistenter Form auf Datenträgern gespeichert. Hierzu werden Datenbanken genutzt, die das strukturelle Fundament für die Speicherung, Verwaltung und Nutzung von Daten bilden (Steiner 2021). Datenbanken stellen eine systematische Datensammlung von zusammengehörigen Daten dar, die aus Sicht des Nutzers miteinander in Verbindung stehen (Vieweg et al. 2012). Sie werden durch ein Datenbankverwaltungssystem organisiert, abgefragt und gesichert werden (Schicker 2017). Abbildung 2 zeigt die beiden Kernkomponenten eines solchen Systems.

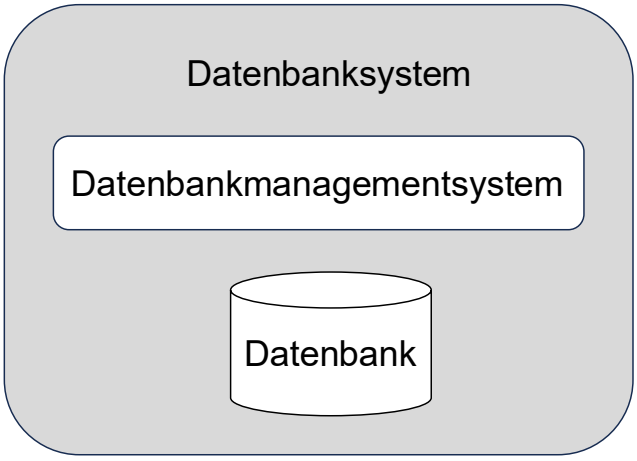

Abbildung 2: Datenbanksystem (eigene Darstellung in Anlehnung an Vieweg et al. (2012))

In Abbildung 2 ist die eigentliche Datenbank zu sehen, in der die Daten physisch gespeichert werden und zum anderen das darüberliegende Softwaremanagementsystem. Dieses sogenannte Datenbankmanagementsystem (DBMS) reguliert den Zugriff, verwaltet Transaktionen

und stellt die Konsistenz der Daten sicher (Mertens et al. 2017). Das DBMS abstrahiert die physische Speicherstruktur und bietet dem Nutzer eine logische Schnittstelle, um Daten effizient zu manipulieren, ohne deren interne Organisation verstehen zu müssen (Vieweg et al. 2012). Diese Abstraktion ermöglicht es, komplexe Operationen wie Joins oder Aggregationen nahtlos auszuführen, während gleichzeitig Redundanzen vermieden und Datensicherheit gewährleistet wird (Wagner et al. 2012). Ein wichtiger Aspekt von Datenbanksystemen ist die Art und Weise, wie sie Daten verwalten und organisieren. Es gibt eine Vielzahl von Datenbankmodellen, wobei es zu jedem Modell unterschiedliche Implementierungen und Hersteller gibt. Die Wahl des Datenbankmodells hängt maßgeblich von den Anforderungen an Skalierbarkeit, Datenstruktur und Konsistenz ab (Meier und Kaufmann 2016).

Relationale Datenbanken organisieren Daten in einer tabellenbasierten Struktur. Die Tabellenspalten werden mit Attributsnamen überschrieben (Meier und Kaufmann 2016). In jeder Zeile ist die Eigenschaft eines einzigen Objektes eingetragen (Vieweg et al. 2012). Ein Datensatz bezeichnet einen einzelnen Eintrag zusammengehöriger Daten in einer Datenbank, der aus einer festen Anzahl von Attributswerten besteht. Die Attribute definieren die Kategorien der gespeicherten Informationen, während die Attributswerte die konkreten Einträge darstellen. Mehrere Tabellen können dabei in einer inhaltlichen Beziehung zueinanderstehen (Vieweg et al. 2012). Die so über Relationen organisierten Daten werden über Primär- und Fremdschlüssel verknüpft (Kemper und Eickler 2015). Ein Primärschlüssel ist ein Feld oder eine Kombination mehrerer Felder, die einen Datensatz eindeutig identifizieren und die Eindeutigkeit sicherstellen als auch die Verwaltung erleichtern. Ein Fremdschlüssel ist ein Attribut, das in einer anderen Tabelle als Primärschlüssel dient und eine Verbindung zwischen den Tabellen herstellt, indem es auf zugehörige Datensätze verweist (Kemper und Eickler 2015). Die Datenstruktur ermöglicht es, Suchen in Datenbanken möglichst schnell zu gestalten (Kleuker 2016). Ein zentrales Konzept in relationalen Datenbanken ist die Transaktion. Diese umfasst eine Reihe von Operationen, durch die die Datenbank ihre Konsistenz bewahrt, während sie von einem definierten Zustand in einen möglicherweise angepassten, jedoch weiterhin konsistenten Zustand übergeht (Kemper und Eickler 2015). Die Transaktion wird entweder vollständig ausgeführt oder vollständig verworfen. Nach Haerder und Reuter (1983) wird dieser Zustand als transaktionskonsistent oder logisch konsistent bezeichnet. Eine Datenbank ist eben nur dann konsistent, wenn sie das Ergebnis von erfolgreichen Transaktionen enthalten. Probleme wie Kommunikationsfehler oder die gleichzeitige Ausführung mehrerer Transaktionen, die zu Zugriffsverletzungen führen kann, stellen Herausforderungen bei Transaktionen dar. Das ACID-Prinzip garantiert, dass Datenbanken selbst bei solchen parallelen Zugriffen oder Systemausfällen in einem validen Zustand bleiben (Haerder und Reuter 1983). Haerder und Reuter (1983) beschreiben das Transaktionsparadigma wie folgt:

Atomicity (Atomarität): Eine Transaktion wird entweder vollständig ausgeführt oder im Fehlerfall vollständig rückgängig gemacht Es existieren keine halbfertigen Zustände.

Consistency (Integritätserhaltung): Es treten keine Regelverstöße oder Dateninkonsistenzen auf. Die Datenbank bleibt vor und nach einer Transaktion in einem konsistenten Zustand.

Isolation (Isolation): Gleichzeitige Transaktionen beeinflussen sich nicht gegenseitig, sodass keine ungewollten Wechselwirkungen entstehen.

Durability (Dauerhaftigkeit): Nach dem erfolgreichen Abschluss einer Transaktion bleiben die Änderungen dauerhaft gespeichert, selbst bei Systemausfällen.

Dieser Mechanismus verhindert inkonsistente Zwischenzustände und unterstreicht die Rolle von Datenbanken als zuverlässige Quelle für unternehmenskritische Daten (Silberschatz et al. 2020). Durch das ACID-Prinzip bieten relationale Datenbanken folglich eine hohe Datenintegrität, stoßen jedoch bei bestimmten Anwendungsfällen an ihre Grenzen. Bei relationalen Datenbanken enthält jede Tabelle eine Menge von Tupeln desselben Typs. Ein Tupel ist eine Menge konkreter Datenwerte, die jeweils eine bestimmte Ausprägung eines Attributs darstellen. Die auf Mengen basierenden Tabellen erlauben nur, dass ein und dasselbe Tupel lediglich einmal vorkommt (Meier und Kaufmann 2016). Dieses Mengenkonzept ermöglicht es, Abfragen und Manipulationen mengenorientiert durchzuführen. Die wichtigste Sprache für solche Operationen ist Structured Query Language (SQL) (Meier und Kaufmann 2016). SQL ist eine deskriptive Sprache, bei der Abfragen allein das gewünschte Ergebnis beschreiben ohne einen konkreten, vollständigen Algorithmus zur Datenabfrage zu erfordern.

Im Gegensatz dazu sind NoSQL-Datenbanken speziell für unstrukturierte Daten und hochskalierbare Anwendungen konzipiert und ergänzen durch NoSQL-Technologien relationale Datenbanken, die hier an ihre Grenzen stoßen. Sie verzichten auf feste Schemata und relationale Strukturen, um Flexibilität und horizontale Skalierbarkeit zu priorisieren (Schicker 2017). Ein Dokumentenmodell speichert Daten in JSON-ähnlichen Objekten, die heterogene Attribute enthalten können (Tiwari 2011). Graphendatenbanken, wie Neo4j, wiederum modellieren Beziehungen explizit, um Netzwerkanalysen zu optimieren (Kemper und Eickler 2015). Der Kompromiss liegt hier jedoch in der Konsistenz. NoSQL-Datenbanksysteme erfüllen die Eigenschaften relationaler Datenbanken nur teilweise und folgen dem CAP-Theorem (Brewer 2000). Oft wird strikte Konsistenz zugunsten von Verfügbarkeit und Ausfalltoleranz geopfert, wodurch eine Transaktion erst nach Verzögerung systemweit sichtbar sein kann. Diese drei Eigenschaften, die nach Brewer (2000) nicht gleichzeitig in einem massiv verteilten System gelten können sind nach Meier und Kaufmann (2016) wie folgt definiert:

Consistency (Konsistenz): Bei Änderungen in einer verteilten Datenbank mit replizierten Knoten sehen alle Lesezugriffe stets den aktuellen Datenstand, unabhängig davon, über welchen Knoten sie erfolgen.

Availability (Verfügbarkeit): Die Anwendung bleibt durchgehend funktionsfähig und liefert innerhalb akzeptabler Zeiten eine Antwort.

Tolerance to network partitions (Ausfalltoleranz): Der Ausfall einzelner Knoten oder Verbindungen beeinträchtigt das Gesamtsystem nicht, und neue Knoten können jederzeit ohne Betriebsunterbrechung hinzugefügt oder entfernt werden.

Datenträger, die zur persistenten Speicherung von Daten genutzt werden, sind oft nicht singulär. Speichersysteme verteilen Daten über mehrere Server oder Speichergeräte hinweg. Sowohl relationale als auch NoSQL-Datenbanken können in sogenannten verteilten Datenbanksystemen (VDBS) eingesetzt werden, um die Leistung zu verbessern, Skalierbarkeit und Ausfallsicherheit zu gewährleisten. Eine zentrale Speicherung der Daten bei Serverproblemen kann sonst zum vollständigen Ausfall führen. Gleichzeitig ermöglichen sie es, geografisch verteilte Nutzer effizienter zu bedienen, ohne auf zeitintensive Methoden wie die ständige Spiegelung aller Daten zurückgreifen zu müssen. Nutzer können von ihrem lokalen Arbeitsplatz aus auf die Daten des gesamten verteilten Systems zugreifen, während die Datenverteilung für sie unsichtbar bleibt und das System im Idealfall wie eine zentrale Datenbank wirkt (Vieweg et al. 2012). In verteilten Datenbanksystemen ist die Allokation entscheidend, da sie festlegt, welche Teile der Datenbank auf welchen Servern gespeichert werden. Dies ermöglicht eine effiziente Nutzung der Ressourcen und eine optimierte Systemleistung (Tanenbaum et al. 2013). Ein weiterer wichtiger Aspekt ist die Fragmentierung, bei der eine Datenbank in kleinere Teile aufgeteilt wird, um sie über mehrere Standorte zu verteilen. Dabei unterscheidet man zwischen horizontaler und vertikaler Fragmentierung. Während bei der horizontalen Fragmentierung die Datensätze einer Tabelle auf verschiedene Server verteilt werden, erfolgt bei der vertikalen Fragmentierung eine Aufteilung nach Spalten, sodass jeder Server nur bestimmte Attribute speichert (Tanenbaum et al. 2013).

Unabhängig davon, ob eine Datenbank zentral oder verteilt organisiert ist, bleibt die Datenqualität ein entscheidender Faktor für die Nutzbarkeit der gespeicherten Informationen. Insbesondere im Kontext von Data Mining sind präzise und konsistente Daten essenziell für aussagekräftige Analysen. Datenqualität ist ein mehrdimensionales Konzept, das beschreibt, inwieweit ein Datenbestand die erforderlichen Merkmale aufweist, um definierte Anforderungen zu erfüllen (Krahl et al. 1998). Im Bereich von Datenbanken und Data Mining geht Datenqualität über technische Faktoren wie Vollständigkeit und Konsistenz hinaus und berücksichtigt auch die Zweckmäßigkeit der Daten, um ihre Verwendbarkeit für spezifische Analysen zu gewährleisten. Die Komplexität des Begriffs Datenqualität zeigt sich in vielfältigen Modellen, wie

(Wang und Strong 1996) in ihrer umfangreichen Untersuchung zum Verständnis des Begriffs darlegen. (English 1999) differenziert zwischen Datendefinitionsqualität, die Vollständigkeit und Granularität umfasst, und Architekturqualität, die Zugriffsfähigkeit und Interpretierbarkeit betont. (Redman 1998) fokussiert auf strukturelle Konsistenz und Korrektheit, während Böttiger et al. (2001) semiotische Ebenen wie syntaktische Einheitlichkeit und pragmatische Zweckeignung ergänzen. Rohweder et al. (2008) systematisieren Dimensionen der Datenqualität in Kategorien wie systemunterstützt und zweckabhängig, wobei sie betonen, dass die Eignung für den Anwendungskontext entscheidend ist.

Mangelhafte Datenqualität birgt erhebliche Risiken. Ineffiziente Entscheidungsprozesse durch fehlerhafte Analysen (Kahn et al. 2016), hohe Betriebskosten durch nachträgliche Datenbereinigung (Redman 1998) und sinkendes Vertrauen in Datenpools (Ryu et al. 2006) sind nur einige Beispiele. Um diesen Herausforderungen zu begegnen, wird im Folgenden für diese Arbeit die Datenqualität über unterschiedliche Dimensionen aus der Literatur festgelegt.

- *Syntaktische Korrektheit*: Daten entsprechen einer vorgegebenen Syntax und haben einen konsistenten Aufbau (Krishna et al. 2023).
- *Angemessene Datenmenge*: Das Datenvolumen ist für die jeweilige Aufgabe geeignet und ist weder zu gering noch zu groß. Die Menge der Daten muss den gestellten Anforderungen der Aufgabe entsprechen (Lee et al. 2006).
- *Relevanz*: Misst, inwieweit Daten in Abhängigkeit vom Kontext für eine bestimmte Aufgabe, Anwendung oder Entscheidungsfindung nützlich, anwendbar und hilfreich sind (Rogova und Bosse 2010).
- *Eignung für den Zweck*: Beurteilt, ob die Daten geeignet sind, eine bestimmte Aktion auszuführen oder eine Anforderung zu erfüllen, insbesondere im Hinblick auf Nutzeraufgaben und -ziele ((Michnik und Lo 2009; Bernhard und Dragan 2007).

Die gelisteten Dimensionen sind entscheidend für die Qualität von Daten und somit für den Erfolg jeder Aufgabe, die auf diesen Daten basiert. Syntaktische Korrektheit ist eine grundlegende Anforderung, insbesondere wenn Daten nur maschinell verarbeitet und analysiert werden können. Dies kann nur gelingen, wenn sie den vorgegebenen Regeln für Format und Aufbau folgen (Krishna et al. 2023). Wenn Daten beispielsweise in einem falschen Datumsformat vorliegen, können Algorithmen diese nicht interpretieren, wodurch wiederum ein Vorverarbeitungsschritt nötig wird (Cai und Zhu 2015). Die angemessene Datenmenge ist relevant, weil sowohl ein Mangel als auch ein Überfluss an Daten problematisch sein kann (Lee et al. 2006). Zu wenig Daten könnten nicht ausreichen, um statistisch valide Schlüsse zu ziehen oder robuste Modelle zu entwickeln (Rogova und Bosse 2010). Eine zu große Menge an irrelevanten Daten hingegen kann die Analyse erschweren und den Rechenaufwand unnötig erhöhen. Dies würde dazu führen, dass wichtige Informationen in einer Flut von unwesentlichem Material

untergeht (Batini und Scannapieco 2016). Es ist daher wichtig, dass genügend Daten vorhanden sind, um die Fragestellungen effizient bearbeiten zu können. Die Relevanz stellt sicher, dass die Daten überhaupt geeignet sind, die Aufgabe zu erfüllen. Wenn Daten formal korrekt und in angebrachter Menge zur Verfügung stehen, jedoch keinen direkten Bezug zum Thema haben, sind sie wertlos und können sogar Irre führend sein (Rogova und Bosse 2010). Schließlich fasst die Eignung für den Zweck die Notwendigkeit zusammen, dass Daten spezifisch für die geplante Aufgabe geeignet sein müssen (Michnik und Lo 2009). Ein Datenbestand mag für eine deskriptive Analyse relevant sein, aber für eine komplexe Modellierung ungeeignet, wenn beispielsweise die Granularität der Daten nicht passt (Bernhard und Dragan 2007).

2.3 Vorgehensmodelle der Wissensentdeckung in Datenbanken

Im Kontext der Wissensentdeckung in Datenbanken haben sich im Laufe der Zeit unterschiedliche Vorgehensmodelle etabliert, die strukturierte Abläufe zur Extraktion verwertbaren Wissens aus Daten definieren (Chu und Yong 2021). Ein Vorgehensmodell setzt sich aus einer Reihe von Prozessphasen zusammen, die als strukturierter Leitfaden für die Durchführung dienen. Sie erleichtern die Planung und Steuerung von Prozessen der Wissensentdeckung in Datenbanken, wodurch sowohl Zeit als auch Kosten gespart werden. Die Hauptziele eines solchen Modells sind, ein tieferes Verständnis für das Projekt und die zugrunde liegenden Daten zu erlangen, die Daten effizient aufzubereiten und zu analysieren sowie die Ergebnisse auszuwerten und anzuwenden (Lieber et al. 2013). Durch die Verwendung von Vorgehensmodellen wird der Prozess der Wissensentdeckung in Datenbanken effizienter und systematischer gestaltet, wodurch wiederum die Qualität und Relevanz des extrahierten Wissens verbessert wird. Ohne strukturierte Vorgehensmodelle droht sonst die Gefahr, irrelevante oder triviale Muster zu produzieren, da unvorbereitete Daten häufig unvollständig, verrauscht oder redundant sind (Runkler 2010).

Entsprechend werden in diesem Kapitel ausgewählte Vorgehensmodelle vertieft und eine anschließende Gegenüberstellung schafft den konzeptionellen Rahmen, um die Rolle der Datenvorverarbeitung im Gesamtprozess der Wissensentdeckung zu verankern.

Cross Industry Standard Process for Data Mining (CRISP-DM)

Das CRISP-DM-Modell nach Chapman et al. (2000) ist ein praxisorientiertes Vorgehensmodell aus dem Jahr 2000. Branchenübergreifend werden Data-Mining-Projekte durch dieses standardisiert werden. Es ist in sechs Phasen unterteilt, die den gesamten Lebenszyklus eines Projekts von der Zieldefinition bis zur Implementierung der Ergebnisse, abdecken (Chapman et al. 2000). Das Modell ist in Abbildung 3 dargestellt.

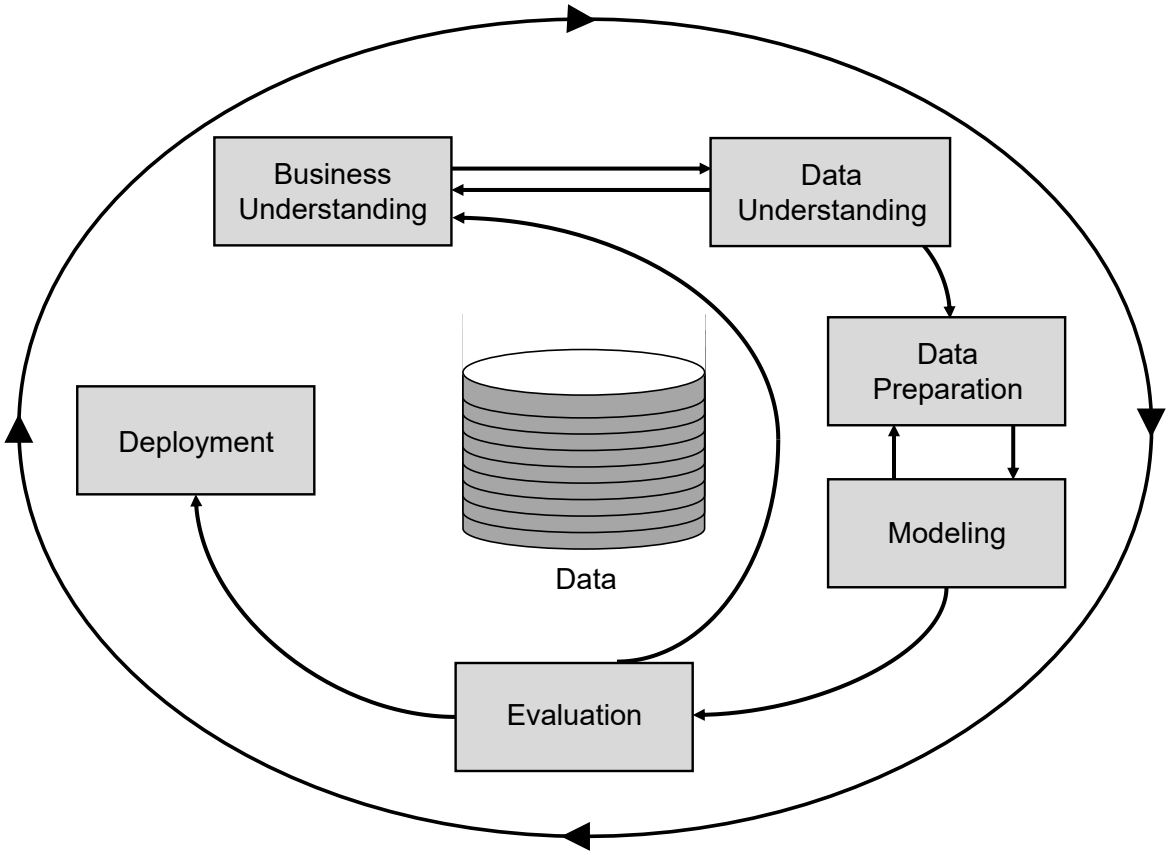

Abbildung 3: CRISP-DM-Vorgehensmodell (eigene Darstellung in Anlehnung an Chapman et al. (2000))

Das in Abbildung 3 zu sehende CRISP-DM-Vorgehensmodell wird im Folgenden Absatz nach Chapman et al. (2000) beschrieben.

1. *Geschäftsverständnis (Business Understanding)*: Die geschäftlichen Ziele und Anforderungen werden analysiert und in eine konkrete Data-Mining-Problemstellung überführt. Dabei fließen Expertenwissen sowie organisatorische Rahmenbedingungen ein. Ziel ist es, die Problemstellung klar zu definieren und eine erste Strategie für das weitere Vorgehen zu entwickeln.

2. *Datenverständnis (Data Understanding)*: Gesammelte Daten werden explorativ analysiert, um erste Muster zu erkennen, Zusammenhänge zu verstehen und potenzielle Qualitätsprobleme wie fehlende Werte oder Inkonsistenzen zu identifizieren. Diese Phase hilft, ein fundiertes Verständnis der Datenbasis zu erlangen und Hypothesen für verborgene Informationen zu formulieren.

3. *Datenaufbereitung (Data Preparation)*: Daten für die Modellierung werden vorbereitet, indem sie bereinigt, transformiert und gegebenenfalls in relevante Teilmengen unterteilt werden. Dabei werden Attribute ausgewählt, Formate angepasst und fehlerhafte oder irrelevante Daten entfernt.

4. *Modellierung (Modeling)*: In dieser Phase werden verschiedene Modellierungsverfahren auf die vorbereiteten Daten angewendet. Die Algorithmen werden ausgewählt und ihre Parameter

optimiert, um möglichst präzise und aussagekräftige Modelle zu erstellen. Falls erforderlich, erfolgen Rückkopplungen zur Datenvorbereitung, um die Datenbasis weiter anzupassen.

5. *Auswertung (Evaluation)*: Die erstellten Modelle werden anhand sowohl technischer als auch geschäftlicher Kriterien validiert. Dabei wird überprüft, ob sie die ursprünglichen Anforderungen erfüllen und sinnvolle Ergebnisse liefern. Falls sich Schwachstellen zeigen, kann eine Anpassung oder erneute Optimierung erforderlich sein. Am Ende dieser Phase wird entschieden, ob und wie die gewonnenen Erkenntnisse genutzt werden.

6. *Einsatz (Deployment)*: Die finale Phase umfasst die praktische Umsetzung der gewonnenen Erkenntnisse. Je nach Anwendungsfall kann dies die Integration der Modelle in bestehende Entscheidungssysteme, die Automatisierung von Prozessen oder die Erstellung von Berichten umfassen. Während die technische Implementierung häufig von Datenanalysten begleitet wird, liegt die eigentliche Anwendung der Ergebnisse oft in den Händen der Nutzer.

CRISP-DM zeichnet sich durch seine zyklische Struktur aus. Der äußere Kreis im Modell symbolisiert, dass Data-Mining-Projekte nie abgeschlossen sind, sondern kontinuierlich optimiert oder auf neue Anwendungsfälle übertragen werden können (Chapman et al. 2000). Diese Eigenschaft spiegelt sich in der detaillierten Modelldokumentation wider, die die Integration generischer und domänenspezifischer Komponenten adressiert und somit universelle Einsetzbarkeit gewährleistet.

Vorgehensmodell nach Fayyad et al. (1996)

Das Vorgehensmodell nach Fayyad et al. (1996), das als grundlegender Ansatz zur Wissensentdeckung in Datenbanken gilt, veranschaulicht. Es besteht aus insgesamt neun Phasen, die in fünf zentrale Kernphasen sowie vier ergänzende Schritte unterteilt sind. Die Kernphasen zeichnen sich durch iterative Rückkopplungen sowie die Einbindung von Domänenwissen aus. Aufgrund ihrer Überlappungen und wechselseitigen Abhängigkeiten ist eine eindeutige Trennung der Phasen nur eingeschränkt möglich (Cleve und Lämmel 2024). Im Folgenden werden nach Fayyad et al. (1996) die neun Phasen dieses Modells detailliert erläutert, wobei das Vorgehensmodell in Abbildung 4 dargestellt ist.

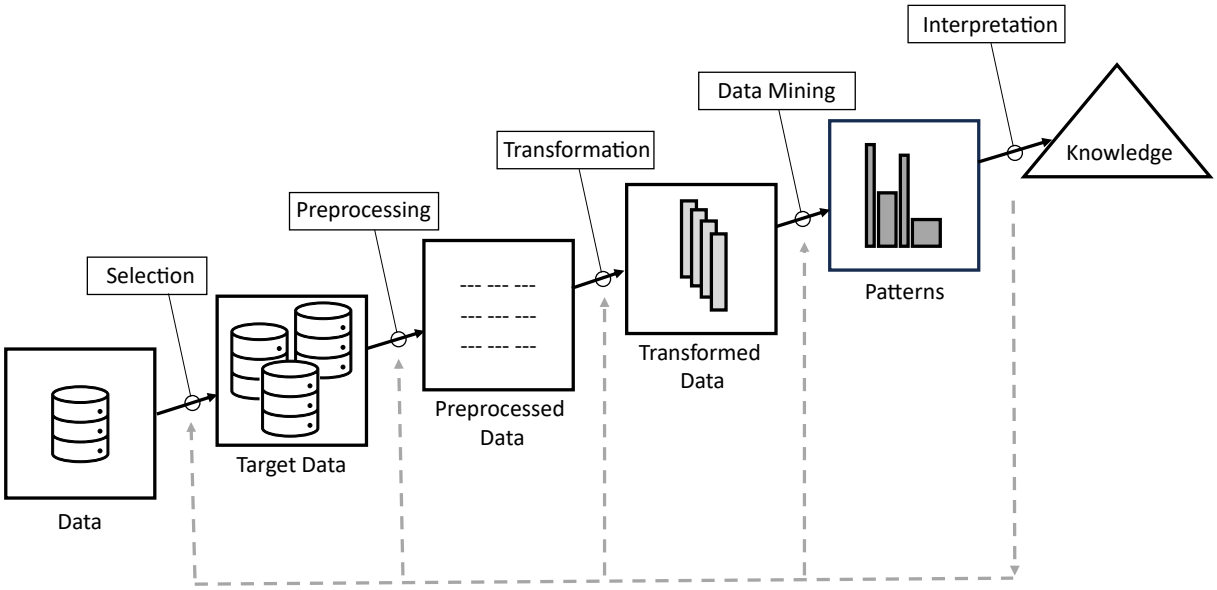

Abbildung 4: Vorgehensmodell nach Fayyad (eigene Darstellung Fayyad et al. (1996)).

Das in Abbildung 4 zu sehende Vorgehensmodell wird im Folgenden Absatz nach Fayyad et al. (1996) beschrieben.

1. *Bereitstellung von Hintergrundwissen*: Der fachliche Kontext wird analysiert und die Projektziele aus Anwendersicht definiert. Hierbei geht es darum, ein tiefgehendes Verständnis für die Branche, Geschäftsprozesse oder spezifische Anwendungsfälle zu entwickeln, um sicherzustellen, dass die Analyse den jeweiligen Anforderungen entspricht.

2. *Erstellung des Zieldatensatz*: Aus der Gesamtmenge der verfügbaren Daten wird ein relevanter Ausschnitt für die Analyse ausgewählt. Dazu werden gezielt Variablen oder Datensegmente identifiziert, die für die Fragestellung essenziell sind.

3. *Datenbereinigung und Datenvorverarbeitung*: Datenmängel wie Rauschen, fehlende Werte oder Inkonsistenzen werden behandelt. Ziel ist es, die Datenqualität zu verbessern, sodass eine verlässliche Analyse möglich wird.

4. *Datenreduktion und Projektion*: Irrelevante oder redundante Informationen werden eliminiert, um die Komplexität der Daten zu verringern. Überflüssige oder weniger relevante Informationen werden entfernt oder in eine kompaktere Form überführt.

5. *Anpassung auf einen Aufgabenbereich des Data Mining*: Die Zielsetzung wird auf einen Aufgabenbereich des Data Mining abgeglichen.

6. *Modellauswahl zur angemessenen Wissensrepräsentation*: In diesem Schritt werden Experimente mit unterschiedlichen Modellen durchgeführt, um die bestmögliche Lösung für den spezifischen Anwendungsfall zu finden.

7. *Data Mining*: In den vorbereiteten Daten werden relevante Muster in einer repräsentativen Darstellungsform entdeckt. Die Qualität der entdeckten Muster hängt maßgeblich davon ab, wie sorgfältig die vorhergehenden Schritte des Prozesses durchgeführt wurden.

8. *Ergebnisinterpretation*: Die Ergebnisse werden, gegebenenfalls mithilfe von Visualisierungen, interpretiert. Falls notwendig, können Anpassungen vorgenommen und frühere Schritte wiederholt werden, um die Modelle weiter zu optimieren.

9. *Operationalisierung des Wissens*: Die gewonnenen Erkenntnisse werden in bestehende Systeme integriert oder für Entscheidungsträger aufbereitet. Ziel ist es, die neu gewonnenen Informationen praxisnah nutzbar zu machen und gegebenenfalls mit bestehendem Wissen abzugleichen.

Das Vorgehensmodel nach Fayyad et al. (1996) beschreibt den Prozess der Wissensentdeckung und betont die enge Verzahnung der strukturierten Abfolge der Phasen. Insbesondere bei unzureichenden Ergebnissen oder Fehlern, erlauben sie eine iterative Rückkopplung. Bei unbefriedigenden Ergebnissen oder zur Korrektur von Fehlern, kann nach jeder Kernphase revidiert und flexibel zurückgesprungen werden.

Weitere Modelle und Gegenüberstellung

Neben den etablierten Modellen nach Fayyad et al. (1996) und CRISP-DM existieren zahlreiche domänenspezifische Anpassungen. Das Knowledge-Discovery-in-Industrial-Databases-Modell von (Lieber et al. 2013) erweitert den Prozess der Wissensentdeckung in Datenbanken um industrielle Anforderungen wie die Erfassung des IT-Ist-Zustands und die Integration von Produktionsdaten aus heterogenen Quellen.

Das SEMMA-Modell des Softwareunternehmens SAS (Wöstmann et al. 2017) konzentriert sich hingegen auf die Modellierungsphase und verzichtet auf geschäftliche Kontextschritte. Der Prozess beginnt mit der Stichprobe *(*Sample*)*, bei der die relevanten Daten ausgewählt werden. Anschließend folgt die Phase des Datenverständnisses (Explore*)*, in den Beziehungen zwischen den einzelnen Attributen analysiert werden. In der darauffolgenden Phase der Datenmodifikation (Modify) werden die Daten transformiert und für die Modellierung vorbereitet. Als nächste Phase erfolgt die Modellierung (Model), in der verschiedene Data-Mining-Methoden angewendet werden. Abschließend werden die Ergebnisse in der Phase Beurteilung (Assess) evaluiert. SEMMA eignet sich besonders für Projekte mit klarem Fokus auf technische Datenanalyse.

Trotz ihrer Unterschiede folgen die verschiedenen Vorgehensmodelle der Wissensentdeckung in Datenbanken grundlegenden Prinzipien. Sie betonen die iterative Natur des Prozesses, unterstreichen Datenqualität und die Notwendigkeit, Domänenexperten einzubeziehen, um identifizierte Muster in verwertbares Wissen zu überführen (Lieber et al. 2013). Unabhängig von

ihrer spezifischen Gestaltung lassen sich vier zentrale Phasen identifizieren, die allen vorgestellten Modellen gemein sind. Zunächst gibt es eine Zieldefinition, bei der die Problemstellung und Analyseziele präzisiert werden. Darauf folgt die Datenvorverarbeitung, die wesentliche Schritte wie die Auswahl relevanter Daten, deren Bereinigung und Transformation umfasst, um eine geeignete Grundlage für die Analyse zu schaffen. Anschließend erfolgt das eigentliche Data Mining, bei dem relevante Zusammenhänge und Strukturen in den Daten aufgedeckt werden. Abschließend werden die gewonnenen Erkenntnisse in einer Ergebnisevaluation geprüft, interpretiert und hinsichtlich ihrer Relevanz für die ursprüngliche Problemstellung bewertet (Lieber et al. 2013).

Als Grundlage dieser Arbeit wird das Vorgehensmodell nach Fayyad et al. (1996) verwendet. In diesem Modell wird ausdrücklich betont, dass nicht nur der eigentliche Data-Mining-Schritt, sondern auch die vorhergehenden Schritte der Datenvorverarbeitung entscheidend für den Erfolg sind. Die Arbeit konzentriert sich auf Verfahren der Datentransformation und Datenreduktion, die nach Fayyad et al. (1996) als eine explizite Phase aufgeführt werden. Während in anderen Vorgehensmodellen wie CRISP-DM die Transformation oft als Teil der breiter gefassten Datenvorbereitung betrachtet wird, ist sie im Modell nach Fayyad et al. (1996) als eigenständige Phase definiert.

Die Wissensentdeckung in Datenbanken geht über die reine Anwendung von Data-Mining-Algorithmen hinaus und umfasst auch vorbereitende sowie nachbereitende Schritte (Fayyad et al. 1996). Innerhalb dieses Prozesses wird Data Mining als ein bedeutender Schritt betrachtet, der dazu beiträgt, versteckte Muster aus den Daten zu extrahieren (Probst et al. 2006). Das folgende Kapitel 2.4 widmet sich entsprechend dem Data Mining und dessen Methoden im Detail.

2.4 Data Mining

Nach dem Vorgehensmodelle nach Fayyad et al. (1996) stellt Data Mining eine zentrale/essentielle Phase dar, die entscheidend für die Extraktion von zuvor unbekanntem und potenziell nützlichem Wissen aus Daten ist (Cleve und Lämmel 2024). Daher ist ein tiefgehendes Verständnis der Ziele, Methoden und Anforderungen des Data Mining unerlässlich, um die Notwendigkeit und spezifischen Aufgaben der Datenvorverarbeitung im Kontext dieses etablierten Vorgehensmodells vollständig zu erfassen und wissenschaftlich zu fundieren.

Trotz der weiten Verbreitung des Begriffs ist dieser in gewisser Hinsicht irreführend. Anders als der Name vermuten lässt, geht es beim Data Mining nicht um das schürfen nach Rohdaten, sondern um die Suche nach Wissen in bereits vorhandenen Datenbeständen. In der Literatur finden sich zahlreiche Definitionen, die unterschiedliche Aspekte betonen. Das Gabler Wirtschaftslexikon beschreibt Data Mining als die Anwendung von Methoden zur automatischen

Extraktion empirischer Zusammenhänge aus Datenbanken, während die Gartner Group (1996) den Fokus auf Mustererkennungstechnologien legt, die große Datenmengen durchsuchen. Linoff und Berry (2011) heben die automatische oder semi-automatische Analyse hervor, um signifikante Muster und Regeln zu identifizieren. Das Hauptziel des Data Mining ist es, implizit vorhandenes Wissen in umfangreichen Datenbeständen zu entdecken und explizit zu machen (Düsing 2006). Hand (2001) betont die Entdeckung überraschender Beziehungen und neuartiger Zusammenfassungen. Thearling (2010) reduziert den Begriff auf die automatisierte Gewinnung von Vorhersageinformationen, was den anwendungsorientierten Nutzen unterstreicht. Gemein ist allen Definitionen der explorative Ansatz. Data Mining durchforstet strukturierte oder unstrukturierte Daten, um verborgene Trends, Korrelationen oder Anomalien aufzudecken. In Anlehnung an Hand (2001), Linoff und Berry (2011) und Thearling (2010) wird für diese Arbeit Data Mining definiert als die Anwendung von Methoden und Algorithmen, um hypothesenfrei Wissen zu extrahieren und vorwiegend unbekannte Muster in Daten zu entdecken. Hierin beschreiben Muster wiederkehrende Strukturen, Trends oder Eigenschaften (Cleve und Lämmel 2024). Im Kontext der Wissensentdeckung geht Data Mining über die Ausführung von SQL-Abfragen hinaus und zielt darauf ab komplexe Zusammenhänge zu erkennen. Eine grundlegende Unterscheidung bei den im Data Mining verwendeten Verfahren ist die zwischen überwachtem Lernen und unüberwachtem Lernen (Cleve und Lämmel 2024). Beim überwachten Lernen werden dem Verfahren Beispiele vorgegeben, bei denen das gewünschte Ergebnis bereits bekannt ist. Diese somit gelabelten Daten werden genutzt, um ein Modell zu trainieren, das dann auf neue, unbekannte Daten angewendet werden kann. Beim unüberwachten Lernen hingegen sind die zu findenden Muster oder Gruppen im Vorhinein nicht bekannt und das Verfahren versucht diese selbstständig in den Daten zu entdecken (Cleve und Lämmel 2024). Nachkommend werden einige Aufgabenbereiche des Data Mining behandelt.

Klassifikationen zielen darauf ab, Objekte anhand ihrer Attribute vordefinierten Klassen zuzuordnen, wobei die Klassenstruktur bereits im Voraus bekannt ist (Leskovec et al. 2020). Ein Anwendungsbeispiel ist die Risikoprüfung von Kunden, bei der Kunden einer Risikoklasse zugeordnet werden (Düsing 2006). Somit handelt es sich hier um überwachtes Lernen, das gelabelte Trainingsdaten benötigt, um Modelle wie Entscheidungsbäume, k-Nearest-Neighbour (k-NN) oder künstliche, neuronale Netze zu trainieren (Cleve und Lämmel 2024). In jedem Fall wird eine diskrete Zielvariable benötigt, die die Klassenzugehörigkeit angibt. Bei Entscheidungsbaumverfahren, wie ID3 oder C4.5 Algorithmen (Han 2012), wird der Datenbestand schrittweise anhand von Merkmalen aufgeteilt, bis homogene Gruppen entstehen (Düsing 2006). Das Ergebnis ist ein Baum, der zur Klassifikation oder zur Generierung von Wenn-Dann-Regeln genutzt werden kann. Bei künstlichen neuronalen Netzen lernt das Netzwerk die Gewichte der Verbindungen von Merkmalswerte, um korrekte Klassen auszugeben (Düsing

2006). Es können numerische Daten verarbeitet werden. Die Güte wird bei neuronalen Netzen und bei Entscheidungsbäumen durch die Fehlklassifikationsquote auf einem unabhängigen Testdatenbestand gemessen (Düsing 2006). Das Verfahren k-NN ist das bekannteste und am häufigsten angewendete Verfahren des Instanz-basierten Lernens (García et al. 2015). Es ist besonders geeignet für kontinuierliche, metrische Attribute, während ordinale und nominale Daten eher ungeeignet sind (Cleve und Lämmel 2024). Um eine Vorhersage für eine neue Instanz zu treffen, wird eine Distanzfunktion verwendet, um die ähnlichsten Trainingsinstanzen zu finden (García et al. 2015). Weitere typische Verfahren sind Support Vector Machines (Han 2012) oder Regel-Lernverfahren wie die Algorithmen RIPPER und Prism (Witten et al. 2017). Kritisch bei Verfahren der Klassifikation ist die Vermeidung von Overfitting, weshalb die Trennung von Trainings- und Testdaten sowie Kreuzvalidierung essenziell sind (Leskovec et al. 2020).

Clusteranalysen gruppieren Objekte in Cluster, deren Mitglieder sich ähneln, während sie sich von anderen Clustern unterscheiden (Jain 2013). Als unüberwachtes Lernverfahren benötigt sie keine vordefinierten Klassen, sondern nutzt Distanzmaße wie die euklidische Distanz oder Dichtebasierten Ansätze zur Strukturierung (Xu und Wunsch 2005). Die Verfahren basieren auf der Berechnung von Abstands- oder Ähnlichkeitsmaßen, wie beispielsweise der Cosinus-Abstand (Cleve und Lämmel 2024). Clusterverfahren können hierarchisch, wie durch das Bilden einer Baumstruktur von Clustern, oder partitionieren, also teilen Daten in eine feste Anzahl von Clustern auf, vorgehen (Düsing 2006). Hierarchische Methoden ermöglichen eine flexible Anpassung an Daten mit unterschiedlicher Form und Größe (Runkler 2010). Die Eignung der Eingabedaten hängt stark von den Datentypen der Attribute ab, da diese die Wahl der Abstands- oder Ähnlichkeitsmaße bestimmen (Han 2012). Numerische Daten sind oft bevorzugt, besonders wenn Distanzmaße verwendet werden (García et al. 2015). Fehlende Daten, Rauschen und Ausreißer können die Leistung beeinträchtigen.

Regression modelliert Beziehungen zwischen Variablen, um kontinuierliche Zielwerte vorherzusagen (Jain 2013). Diese Aufgabe ist überwacht und zielt auf die Vorhersage oder Schätzung des Wertes einer kontinuierlichen Zielvariable ab (Düsing 2006). Sie basiert auf statistischen Verfahren wie linearer oder polynomialer Regression und berücksichtigt dabei die Verteilung der Merkmale, sei es univariat oder multivariat. Verfahren wie Entscheidungsbäume und künstliche neuronale Netze können für Regressionsaufgaben eingesetzt werden. Es muss eine kontinuierliche, quantitative Zielvariable vorhanden sein (Düsing 2006). Die Wahl der Methode hängt von der Datenstruktur und der Fragestellung ab, wobei die Interpretierbarkeit des Modells oft gegen seine Vorhersagegenauigkeit abgewogen wird (Hastie et al. 2009).

Wirkzusammenhänge analysieren kausale Beziehungen zwischen Variablen, etwa wie bestimmte Faktoren Ereignisse beeinflussen. Diese unüberwachte Aufgabe konzentriert sich

also auf die Identifizierung von Beziehungen oder Zusammenhängen zwischen Merkmalen oder deren Ausprägungen in den Daten. Deren Ergebnisse können teilweise zu Prognosemodellen erweitert werden (Düsing 2006). Im Gegensatz zu Korrelationen implizieren sie eine Richtung der Beeinflussung, die jedoch statistisch schwer zu validieren ist (Küppers 1999). Verfahren wie Assoziationsregeln oder Bayes'sche Netze versuchen, solche Zusammenhänge in homogenen Datengruppen abzuleiten, stoßen jedoch oft an Grenzen, da Kausalität experimentelle Kontrolle erfordert (Cleve und Lämmel 2024). Diese Verfahren laufen in zwei Phasen ab. Zunächst dem Auffinden häufiger Itemsets und anschließend den zugehörigen Regeln (Düsing 2006). Hinsichtlich der Anforderungen an die Eingabedaten, sind Daten notwendig, bei denen das gemeinsame Auftreten von Attributen von Interesse ist, wie beispielsweise Transaktionsdaten (Düsing 2006). Regel-Lernverfahren und Assoziationsregelverfahren erfordern oft nominale oder diskretisierte Daten (García et al. 2015). Die Relevanz von Assoziationsregeln wird typischerweise mit der Häufigkeit des gemeinsamen Auftretens (Support) und der bedingten Wahrscheinlichkeit bewertet (Han 2012).

Data Mining hat das Ziel, Wissen aus Daten zu extrahieren (Cleve und Lämmel 2024). Die Notwendigkeit der Datenvorverarbeitung ergibt sich hierbei aus dem Zustand der realen Daten und den spezifischen Anforderungen der Data-Mining-Algorithmen (García et al. 2016). Reale Daten, wie sie beispielsweise in Unternehmensdatenbanken anfallen, sind typischerweise unvollkommen und enthalten Inkonsistenzen und Redundanzen (García et al. 2016). Die enormen Mengen an gesammelten Daten, die schnell wachsen, übersteigen die menschliche Fähigkeit zur Verarbeitung ohne leistungsstarke Werkzeuge (Fayyad et al. 1996). Dies führt zu einer Situation, die Han (2012) als „data rich but information poor" beschreiben. Gleichzeitig haben die unterschiedlichen Data-Mining-Algorithmen spezifische Anforderungen an die Eingabedaten. Sie benötigen Daten in einer bestimmten Menge, Struktur und in einem bestimmten Format, um effektiv arbeiten zu können (García et al. 2016). Bestimmte Verfahren arbeiten entweder nur oder bevorzugt mit einem bestimmten Datentyp. So eignen sich beispielsweise für das k-NN kontinuierliche, metrische Attribute besonders gut (Cleve und Lämmel 2024). Die Effektivität von Data-Mining-Algorithmen ist unmittelbar an die Qualität der Eingabedaten gebunden, was die Datenvorverarbeitung zu einer unverzichtbaren Vorstufe macht (Han 2012). Sie ist in der Lage, die Daten an die Anforderungen anzupassen, die jeder Data-Mining-Algorithmus stellt und ermöglicht so die Verarbeitung von Daten, die sonst unmöglich wäre (García et al. 2016). Selbst die fortschrittlichsten Algorithmen scheitern, wenn die Daten inkonsistent, verrauscht oder unvollständig sind. Mit entsprechend nicht vorverarbeiteten Daten, kann der Data-Mining-Algorithmus möglicherweise nicht ausgeführt werden, Fehler zur Laufzeit melden oder Ergebnisse liefern, die keinen Sinn ergeben oder nicht als präzises Wissen betrachtet werden (García et al. 2016). Vor diesem Hintergrund wird in Kapitel 2.5 die Datenvorverarbeitung untersucht.

2.5 Datenvorverarbeitung

Die Gegenüberstellung verschiedener Vorgehensmodelle zeigt, dass die Datenvorverarbeitung in allen Modellen eine essenzielle Phase darstellt. Sie bildet die Brücke zwischen rohen und analysereifen Daten. In der Praxis liegen Daten selten in einem optimalen Zustand vor. Häufig sind sie fragmentiert, verrauscht oder inkonsistent, wodurch die Leistung von Algorithmen beeinträchtigt wird (Little und Rubin 2019). Ziel der Datenvorverarbeitung ist es, Daten in einen Zustand zu überführen, der sowohl technisch kompatibel mit den Anforderungen der Algorithmen des nachgelagerten Data-Mining-Schrittes als auch inhaltlich aussagekräftig ist, um relevante Muster klar abzubilden (García et al. 2015). Die Vorverarbeitung wird verwendet, um einen Datenbestand für die Verarbeitung durch Data-Mining-Algorithmen vorzubereiten (Ali und Faraj 2014).

In Kapitel 2.3 wird das Vorgehensmodell nach Fayyad et al. (1996) als methodische Grundlage für diese Arbeit herangezogen. Demnach befasst sich dieses Kapitel mit der dritten Phase zur Datenbereinigung und Datenvorverarbeitung. Hierzu ist anzumerken, dass Fayyad et al. (1996) von einem bereits integrierten Datenbestand ausgeht. Während moderne Datensätze oft in strukturierten Formaten wie relationalen Datenbanken gespeichert werden und vermehrt automatisiert entstehen, wie etwa durch sensorbasierte Erfassung, entstehen in Unternehmen nichtsdestotrotz häufig Datenbestände aus heterogenen Quellen. Dies führt zu strukturellen und semantischen Inkonsistenzen mit dessen Lösung sich die Datenintegration auseinandersetzt (Gheware et al. 2014). In diesem Zusammenhang werden im Folgenden die maßgeblichen Aufgaben der dritten Phase nach dem Vorgehensmodell von Fayyad et al. (1996) und ihre funktionale Bedeutung im Gesamtprozess eingehend erläutert. Der Vollständigkeit wegen wird zusätzlich zunächst das Gebiet der Datenintegration beleuchtet.

Datenintegration

Die Datenintegration zielt darauf ab, vereinheitlichten Zugriff auf Daten zu ermöglichen, die in mehreren, autonomen Quellen gespeichert sind (Batini und Scannapieca 2006), während Redundanzen und Inkonsistenzen vermieden werden sollen (García et al. 2015). Dies schafft die Grundlage für automatisierte oder teilautomatisierte Prozesse der Datenanalyse. Die Notwendigkeit hierfür ergibt sich dadurch, dass Daten über verschiedene Systeme und Organisationen hinweg geteilt und integriert werden müssen (Doan 2012). Auf diese Weise können komplexe Abfragen beantworten werden, wie beispielsweise die Bereitstellung einer Kundenübersicht aus verschiedenen Abteilungen (Doan 2012). Herausforderungen ergeben sich hier unter anderem aus systemtechnischen, logischen, administrativen und qualitativen Unterschieden zwischen den Datenquellen (Doan 2012).

Die größte Herausforderung bei der Datenintegration stellt somit die Heterogenität der Daten dar (Doan 2012). Hinzu kommt die Problematik der Datenqualität. Werden Datenbestände mit

geringer Qualität kombiniert, so kann sich das Problem zusätzlich verschärfen (Abdallah et al. 2020). Eine Aufgabe der Datenintegration ist die Schemaintegration, die sich mit der semantischen Ambiguität und der strukturellen Vielfalt befasst (Dong und Srivastava 2015). Ziel ist es eine vereinheitlichte Sicht über die verschiedenen Quellen zu bieten und umfasst Techniken wie Auflösen von Namenskonflikten, Abgleichen von Attributen oder semantisches Mapping (Abdallah et al. 2020). Darüber hinaus ist es oft schwierig zu identifizieren, welche Entitäten aus unterschiedlichen Datensätzen dasselbe Objekt darstellen. Dies wird als Ambiguität der Instanzrepräsentation bezeichnet (Dong und Srivastava 2015). Eine weitere Aufgabe ist folglich das identifizieren und auflösen von Datensätzen, die dasselbe Objekt referenzieren, auch wenn diese unterschiedlich dargestellt werden (Dong und Srivastava 2015). Der Prozess, der sich damit befasst wird als Record Linkage bezeichnet. Dies ist selbst dann noch herausfordernd und notwendig, wenn die Schemata bereits abgeglichen wurden und identisch sind. Die Datenintegration erfolgt auf Datenebene und befasst sich mit Problemen, die unmittelbar die Daten selbst betreffen, wie etwa das Vorhandensein von Duplikaten (Dong und Srivastava 2015). Ein weiteres Problem auf Datenebene ist der Umgang mit Inkonsistenzen. Die Aufgabe der Datenfusion besteht in diesem Zusammenhang darin, bei widersprüchlichen Werten aus verschiedenen Quellen zu entscheiden, welcher Wert in die integrierten Daten übernommen wird (Dong und Srivastava 2015).

Datenbereinigung

Fayyad et al. (1996) verstehen die Datenbereinigung als einen essenziellen, frühen Schritt im iterativen Vorgehensmodell. Sie umfasst drei zentrale Herausforderungen, die oft eng miteinander verbunden sind. Den Umgang mit fehlenden Werten, die Reduktion von Rauschen (Fayyad et al. 1996) und die Identifikation sowie Behandlung von Ausreißern (García et al. 2015). Die Datenbereinigung hat das Ziel, die Genauigkeit und damit die Verwendbarkeit vorhandener Daten zu verbessern, indem fehlerhafte Daten erkannt und korrigiert werden (Hosseinzadeh et al. 2023). Die Behandlung mehrerer Fehlertypen stellt sich als anspruchsvolle Aufgabe heraus, insbesondere weil häufig nicht genügend Informationen vorliegen, um die jeweils richtigen Änderungen zuverlässig vorzunehmen. Mitunter werden betroffene Datensätze gelöscht, was jedoch mit potenziellem Informationsverlust einhergeht (Müller und Freytag 2003). Aus diesem Grund ist oft Domänenwissen von Fachexperten notwendig. Ein Aufwand der sowohl kostspielig als auch zeitaufwendig ist (Müller und Freytag 2003).

Fehlende Werte sind Werte für ein Attribut, die bei der Aufzeichnung nicht eingefügt wurden oder verloren gegangen sind. Sie können durch manuelle Dateneingabefehler, technische Fehler oder systematische Lücken in Erfassungsprozessen entstehen (Chen et al. 2012). Eine einfache Methode zum Umgang mit fehlenden Werten ist das Verwerfen von Datensätzen, die fehlende Werte enthalten. Dies ist jedoch nur praktikabel, wenn die Anzahl der Datensätze mit

fehlenden Werten klein ist (García et al. 2015). Während alternative einfache Methoden wie die Ersetzung durch globale Konstanten eine schnelle Lösung bieten, besteht die Gefahr verzerrter Stichproben, insbesondere wenn die Verteilung der fehlenden Werte nicht zufällig ist (Hippner et al. 2004). Darunter existieren Ansätze wie Maximum Likelihood-Imputation, Imputation basierend auf Fuzzy k-Means Clustering oder Singular Value Decomposition Imputation (García et al. 2012).

Rauschen tritt in Form von zufälliger oder systematischer Fehler in gemessenen Variablen auf und lässt sich in Attributrauschen und Klassenrauschen unterteilen (García et al. 2015). Der Umgang mit Rauschen umfasst einerseits dessen Identifizierung und andererseits Maßnahmen zu dessen Behandlung, die meist durch Filterung erfolgen (García et al. 2015). Zur Detektion von Rauschen eigenen sich grundlegende statistische und deskriptive Verfahren wie beispielsweise Streudiagramme (García et al. 2015). Im Fall numerischer Daten kann die multiple lineare Regression verwendet werden, um Trends in den Attributwerten zu modellieren und so eine Grundlage für Rauschschätzung schaffen. Nach der Identifikation des Rauschens erfolgt ein Korrekturschritt. Bei Attributrauschen spielen neben den stochastischen Abhängigkeiten auch die zugrunde liegende Wahrscheinlichkeitsverteilung des Rauschens eine bedeutende Rolle für die Auswahl geeigneter Korrekturmethoden (García et al. 2015). Die Wahl der geeigneten Methode hängt stark vom jeweiligen Anwendungsbereich ab. Ausreißer, also extrem abweichende Datenpunkte, erfordern eine differenzierte Betrachtung. Sie werden häufig als Störfaktoren interpretiert. Je nach Kontext können sie jedoch wertvolle Hinweise auf kritische Anomalien liefern (Aggarwal 2014). Beim Klassenrauschen ist entscheidend, mit welcher bedingten Wahrscheinlichkeit die beobachtete Klasse von der tatsächlichen Klasse abweicht. Dieses Verhältnis lässt sich oft mithilfe einer Übergangsmatrix darstellen und ist grundlegend, um das Ausmaß des Rauschens zu ermitteln und ein geeignetes Korrekturverfahren auszuwählen (García et al. 2015).

Als Ausreißer werden Datenpunkte bezeichnet, die deutlich von den allgemeinen Merkmalen oder typischen Mustern eines gegebenen Datenbestandes abweichen und aufgrund dessen ein hohes Potenzial haben, fehlerhaft zu sein (Müller und Freytag 2003). Durch die Beziehungen zwischen zwei oder mehr Attributen, stellen Ausreißer komplexe Fehler dar, die durch Integritätsbedingungen nur schwer zu erkennen und beschreiben sind (Müller und Freytag 2003). Statistische Methoden können zur Identifizierung von Ausreißern verwendet werden (García et al. 2015). Je nach Datenstruktur und Anwendungsfall kommen verschiedene Modelle zum Einsatz, darunter proximitätsbasierte, dichtebasierte, statistische und unterraumbasierte Ansätze (Aggarwal 2017).

3 Datentransformation und Datenreduktion

Ein zentrales Problem besteht darin, sogenannte Low-Level-Daten in kompaktere, abstraktere oder besser nutzbare Formen zu überführen (Fayyad et al. 1996). Genau hier setzt die vierte Phase der Transformation des Vorgehensmodells nach Fayyad et al. (1996) an, mit der sich dieses Kapitel beschäftigt. Die Transformation baut auf die zuvor abgeschlossene Phase der Datenvorverarbeitung und Datenbereinigung auf. Hierin finden sich die Gebiete der Datentransformation und Datenreduktion wieder, die die Aufgabe haben, Daten aus den vorangegangenen Phasen in Formen zu überführen, die für die Anwendung von Data-Mining-Algorithmen geeignet sind. Laut Fayyad et al. (1996) zielt diese Phase darauf ab, entweder die Anzahl der relevanten Variablen zu verringern oder eine Darstellung der Daten zu finden, die gegenüber bestimmten Veränderungen beständig bleibt. Das Ziel besteht also darin, die Daten so aufzubereiten, dass der anschließende Data-Mining-Schritt effizient durchgeführt werden kann. Im besten Fall mit einer ebenso hohen Qualität der Ergebnisse wie bei der Nutzung der Originaldaten. Die Datentransformation und Datenreduktion bereiten somit die Daten strukturell und inhaltlich so auf, dass sie für Verfahren der Mustererkennung in Data-Mining-Prozessen geeignet sind und schaffen die Grundlage für eine erfolgreiche Wissensentdeckung.

In Folgenden werden zunächst jeweils für die Datentransformation und die Datenreduktion grundlegende Ziele und Aufgaben erläutert. Anschließend werden konkrete Verfahren vorgestellt, die sich in den gesetzten Aufgaben untergliedern lassen.

3.1 Grundlagen der Datentransformation

Die Datentransformation überführt Rohdaten durch inhaltliche und strukturelle Anpassungen in ein Format, das für die beabsichtigten Analysen geeignet ist und für Data-Mining-Verfahren nutzbar ist (Borrohou et al. 2025). Ohne Transformation können Rohdaten nur schwer effektiv analysiert werden, was zu ungenauen oder irreführenden Schlussfolgerungen führen kann (Borrohou et al. 2025).

In dieser Arbeit folgt die Definition der Datentransformation dem Verständnis von Cleve und Lämmel (2024), die darunter sämtliche Aufgaben verstehen, die darauf abzielen, die inhaltliche Struktur oder Repräsentation von Daten anzupassen, um sie für nachgelagerte Data-Mining-Verfahren nutzbar zu machen. Im Gegensatz zur Datenreduktion, die primär auf die Verringerung der Datenmenge fokussiert (Aggarwal 2014), umfasst die Datentransformation hier Verfahren, die semantische, algorithmische oder technische Kompatibilität herstellen, ohne zwangsläufig die Datenmenge zu reduzieren. Die Aufteilung der Notwendigkeit zur Datentransformation auf Daten- und Informationsebene, führt direkt zur Unterscheidung der Datentransformation in verfahrensunabhängige und verfahrensspezifische Transformationen.

- Die verfahrensspezifischen Transformationen werden angewendet, um die Daten für einen bestimmtes Data-Mining-Verfahren passfähig zu machen. Sie finden in der Regel auf Datenebene statt und deren Notwendigkeit ergibt sich direkt aus der angestrebten algorithmischen Kompatibilität. Beispiele hierfür sind die Diskretisierung numerischer Daten oder auch komplexere Transformationen wie die Fourier-Transformation für Zeitreihendaten, um Datenstrukturen für bestimmte Algorithmen aufzubereiten (Cleve und Lämmel 2024). Diese Transformationen sind also direkt an die Anforderungen der gewählten Analysemethoden geknüpft.
- Die verfahrensunabhängigen Transformationen zielen in erster Linie auf die Informationsoptimierung ab. Sie umfassen Aufgaben, die den Datenbestand im Allgemeinen verbessern, indem Daten verständlicher gemacht und relevante Informationen hervorgehoben werden. Beispiele hierfür sind die Anpassung von Datentypen, das Kombinieren oder Separieren von Attributen oder die Berechnung abgeleiteter Werte (Cleve und Lämmel 2024). Diese Schritte stellen einen in sich konsistenten Datenbestand her und erhöhen die allgemeine Datenqualität, was gerade unabhängig vom später genutzten Data-Mining-Algorithmus von Vorteil ist (Cleve und Lämmel 2024).

3.2 Verfahren der Datentransformation

Wie in Kapitel 3.1 erläutert, gliedern sich die Aufgaben der Transformation in verfahrensspezifische und verfahrensunabhängige Transformationen. In diesem Abschnitt werden entsprechende Verfahren aus der Literatur vorgestellt.

Verfahrensspezifische Transformationen

Verfahrensspezifische Transformationen sind notwendig, um die Kompatibilität mit bestimmten Data-Mining-Algorithmen herzustellen.

Normalisierung ist ein Prozess, bei dem Werte in einen bestimmten Wertebereich skaliert werden (Cleve und Lämmel 2024). Dies ist häufig erforderlich, um einen Datenbestand insbesondere für solche Data-Mining-Verfahren anzupassen, die Distanzmetriken verwenden, wie beispielsweise der k-NN-Algorithmus (Cleve und Lämmel 2024). Das Hauptproblem im Rohdatenbestand, das von Verfahren der Normalisierung adressiert wird, sind die großen Unterschiede in den Bereichen verschiedener Merkmale. Merkmale, die sehr unterschiedliche Skalen haben, können einige Algorithmen durch Merkmale mit größeren Werten unverhältnismäßig stark beeinflusst werden (Ali und Faraj 2014). Die Normalisierung hilft folglich dabei, diese Bereiche anzugleichen. Bei komplexen Messungen können auch systematische Variationen oder technisches Rauschen enthalten (Liu et al. 2019). Hierzu gehören Unterschiede in der Skalierung, dem Offset (Lima und Souza 2023), den Maßeinheiten (Chakraborty und Yeh 2007), den experimentellen Bedingungen (Schmidt et al. 2011), den Messplattformen (Rudy

und Valafar 2011), sowie generelles Rauschen und Inkonsistenzen bei der Probenvorbereitung (Vu et al. 2018). Sie kann auch für die lineare Ordnung in der mehrdimensionalen Statistik genutzt werden, indem Merkmale ihrer ursprünglichen Maßeinheiten enthoben werden und ihr Wertebereich vereinheitlicht wird, um sie vergleichbar zu machen (Nasser et al. 2019). Die Min-Max-Normalisierung transformiert Daten linear in einen Zahlenbereich, beispielsweise zwischen 0 und 1, wobei der kleinste Wert auf 0 und der höchste Wert auf 1 skaliert wird (Cleve und Lämmel 2024). Bei der Z-Wert-Normalisierung wird ein Wert normalisiert, indem die Differenz zwischen diesem Wert und dem Mittelwert durch die Standardabweichung dividiert wird (Cleve und Lämmel 2024). Janssen und Laatz (1994) und Sharma (2022) setzen dieses Verfahren mit der Standardisierung gleich. Insbesondere ist die Z-Wert-Normalisierung nützlich, wenn die Daten einer Normalverteilung folgen (Ali und Faraj 2014). Sie zielt darauf ab, einen Datensatz mit einem Mittelwert von 0 und einer Standardabweichung von 1 zu erstellen. Eine weitere Methode ist die Dezimalskalierung, eine lineare Skalierung, die durch das Verschieben des Dezimalkommas erfolgt. Daten werden so in ein vorgegebenes Intervall wie [0 … 1] transformiert (Cleve und Lämmel 2024). Verfahren der Normalisierung können evaluiert werden, indem ihre Auswirkungen auf die Genauigkeit verschiedener Algorithmen gemessen werden. Beispielsweise das 1-Nearest-Neighbour-Verfahren mit Euklidischer Distanz und Dynamic Time Warping (Lima und Souza 2023). Andere Studien untersuchen den Einfluss der Normalisierung auf das Ergebnis der linearen Ordnung für den Vergleich (Nasser et al. 2019). Dies geschieht durch die Berechnung des Korrelationsgrades zwischen den Ergebnissen verschiedener Verfahren und durch die Darstellung der Anzahl der Objekte, deren Rang sich geändert hat, verglichen mit einer Referenzmethode gemessen (Nasser et al. 2019). (Dillies et al. 2013) führt qualitative Vergleiche von unterschiedlichen Verfahren der Normalisierung anhand von Darstellungen wie Boxplots der gezählten Werte durch, um die Verteilung und Variabilität innerhalb von Gruppen zu beurteilen. Der Variationskoeffizient kann als Metrik für die Variabilität innerhalb von Gruppen verwendet werden. Hierzu ist in Anhang A eine Liste statistischer Kennzahlen zu finden. In der Studie nach (Fujita et al. 2006) werden Vergleiche basierend auf der Robustheit gegenüber Ausreißern durchgeführt. Es werden künstliche Datensätze mit absichtlich eingefügten Ausreißern verwendet. Die Leistung wird anhand des mittleren quadratischen Fehlers zwischen der Regressionskurve des Normalisierungsverfahrens und der bekannten tatsächlichen Kurve bewertet. Statistische Tests wie Wilcoxon und Kolmogrov-Smirnov werden verwendet, um signifikante Unterschiede zu erkennen. In der Literatur basieren Vergleichsstudien oft auf experimentellen Bewertungen mit Benchmark-Datensätzen, die real oder simuliert sind (Lima und Souza 2023; Vu et al. 2018; Rudy und Valafar 2011; Fujita et al. 2006; Dillies et al. 2013).

Skalierung und Normalisierung werden in der Literatur oft gemeinsam oder synonym behandelt (Han 2012). Verfahren der Skalierung dienen dazu, die Wertebereiche von Datensätze

anzugleichen (Cleve und Lämmel 2024). Cleve und Lämmel (2024) verstehen die Skalierung als einen allgemeineren Begriff, der die Transformation von Werten auf eine beliebige Skala beschreibt. Das bestätigt auch Ahsan et al. (2021) und beschreibt Skalierung als allgemeinen Überbegriff für die Transformation von Werten auf eine Skala, sodass sie für Analysezwecke vorbereitet werden. Nach (Fahrmeir et al. 2016) lassen sich Skalenniveaus in hierarchisch in Nominalskala, Ordinalskala, Intervallskala und Verhältnisskala unterteilen. Höhere Skalenniveaus beinhalten die Eigenschaften niedrigerer Stufen, wodurch der Informationsgehalt ansteigt (Fahrmeir et al. 2016). Ein Wechsel zwischen diesen Ebenen ist nur in einer Richtung möglich. Während sich eine Verhältnisskala auf eine Ordinalskala reduzieren lässt, ist die Umwandlung einer Ordinalskala in eine Verhältnisskala nicht ohne willkürliche Annahmen durchführbar (Petersohn 2005).

Standardisierung wird von Gal und Rubinfeld (2018) umfassender diskutiert und als konzeptioneller Rahmen für Kohärenz und Kompatibilität von Datenstrukturen über Systeme hinweg verstanden. Hierin werden Standards für Attribute, Terminologie, Struktur und Organisation von Datensätzen festgelegt mit dem zentralen Ziel die Nutzung, Portabilität und Interoperabilität von Daten zu verbessern (Gal und Rubinfeld 2018). Im mathematischen Sinn, wird die Standardisierung mit der Z-Wert-Normalisierung gleichgesetzt und meint eine Rechenoperation auf numerischen Daten (Sharma 2022). Entsprechend zielt die Standardisierung ebenfalls darauf ab den Datenbestand auf einen Mittelwert von 0 und eine Standardabweichung von 1 zu transformieren (Ali und Faraj 2014). Ein Standardformat, das für viele Algorithmen vorteilhaft ist, besonders wenn die Daten normalverteilt sind.

In dieser Arbeit werden die Aufgaben der Normalisierung, Skalierung und Standardisierung zu dem Bereich linearer Transformationen zusammengefasst.

Verteilungstransformationen werden an dieser Stelle als Gruppierung verschiedener Datentransformationsaufgaben definiert. Hierzu zählen Transformationen zur Anpassung der Datenverteilung. Insbesondere logarithmische Transformationen, Wurzeltransformation, Box-Cox-Transformationen sowie Exponentialtransformationen können dazu gezählt werden. Verteilungstransformationen sind notwendig, wenn Daten nicht den statistischen Voraussetzungen entsprechen, wie etwa fehlende Linearität oder unvergleichbare Skalenniveaus (Janssen und Laatz 1994). Die Box-Cox-Transformation fördert Varianzhomogenität und erleichtert dadurch interpretierbare und verlässliche Analysen (Atkinson et al. 2021). Als parametrische Transformation kann sie besonders stark durch Ausreißer beeinflusst werden (Atkinson et al. 2021). Die logarithmische Transformation wie beispielsweise der Logarithmus zur Basis 10 oder der natürliche Logarithmus können für die Anpassungen an statistische Annahmen genutzt wer-

den (Janssen und Laatz 1994). Auch die Wurzeltransformation sowie die Exponentialtransformation zählen zu typischen arithmetischen Umformungen zur Datenanpassung (Janssen und Laatz 1994).

Diskretisierung dient dazu, die Wertebereiche von numerischen, kontinuierlichen Merkmalen in eine endliche Anzahl von diskretisierten, disjunkten Intervallen zusammenzufassen oder zu partitionieren (Han 2012). Kurz gesagt, wandelt die Diskretisierung numerische Daten in kategoriale um, indem Werte bestimmten Intervallen zugeordnet werden. Ziel ist es, eine kompakte Darstellung zu erzeugen, bei denen möglichst viele Informationen des ursprünglichen Datenbestandes erhalten bleiben (Ramírez-Gallego et al. 2016). Angesichts der Reduzierung des Wertebereiches numerischer Daten auf eine begrenzte Anzahl kategorialer Werte wird die Diskretisierung von (Ramírez-Gallego et al. 2016) als eine Aufgabe der Datenreduktion eingestuft. Es gibt Data-Mining-Algorithmen, die nicht mit metrischen Daten umgehen können und daher diskretisierte Daten verlangen (Cleve und Lämmel 2024). Beispielsweise können die Algorithmen ID3 und Naive Bayes nicht mit metrischen Daten arbeiten und setzen nominale oder ordinale Merkmale voraus, während der k-NN-Algorithmus metrische Daten erfordert (Cleve und Lämmel 2024). C4.5 und Apriori sind weitere Beispiele für Methoden, die in der einen oder anderen Form eine Diskretisierung der Daten erfordern (Ramírez-Gallego et al. 2016). Es gibt verschiedene Verfahren der Diskretisierung, die sich beispielsweise danach unterscheiden, ob sie Klasseninformationen nutzen, wie überwachte oder unüberwachte Diskretisierung, oder welche Richtung sie einschlagen (Han 2012). Die Kodierung diskreter Werte kann selbst dann zu unterschiedlichen Ergebnisse bei abstandsbasierten Algorithmen wie k-NN führen, wenn die ursprüngliche Ordnung erhalten bleibt (Cleve und Lämmel 2024).

Binärisation oder Binärkodierung wird als eine Methode der Umwandlung von Attributwerten aufgeführt (Cleve und Lämmel 2024). Diese Technik wird angewendet, um nicht-numerische oder ordinale Daten in ein binäres Format zu überführen (Cleve und Lämmel 2024). Dies ist insbesondere dann erforderlich, wenn der nachgelagerte Data-Mining-Algorithmus numerische oder metrische Daten voraussetzt, wie beispielsweise neuronale Netze oder der k-NN-Algorithmus (Cleve und Lämmel 2024). Ein gängiges Verfahren der Binärkodierung ist für nominale Attribute das One-Hot-Encoding (Cleve und Lämmel 2024). Für jede Ausprägung eines nominalen Attributs wird ein neues binäres Attribut erzeugt, das den Wert 1 annimmt, wenn diese Ausprägung im Datensatz vorkommt und andernfalls den Wert 0 (Cleve und Lämmel 2024). Diese Umwandlung ist eine strukturelle Anpassung auf Datenebene, da sie die ursprünglichen Attribute durch eine Menge neuer, binärer Attribute ersetzt (Cleve und Lämmel 2024). Fernández et al. (2013) beschreibt Binärisation als Methode, um mehrklassige, unausgeglichene Klassifikationsprobleme in mehrere binäre Teilprobleme zu zerlegen.

Verfahrensunabhängigen Transformationen

Verfahrensunabhängige Transformationen zeichnen sich dadurch aus, dass sie unabhängig vom anschließend eingesetzten Data-Mining-Algorithmus durchgeführt werden können. Diese Transformationen dienen vor allem der inhaltlichen Verdichtung.

Aggregation ist eine Aufgabe der Datentransformation, die Daten auf einer höheren Ebene zusammenfasst und verdichtet (García et al. 2015). Hierbei lassen sich aus den Werten mehrerer Variablen neue, zusammenfassende Kennzahlen für einzelne Fälle ableiten (Janssen und Laatz 1994). Ein einfaches Beispiel ist die Berechnung eines Verlustes oder Überschusses als Differenz zwischen Einnahmen und Ausgaben (Janssen und Laatz 1994). Ebenso die Berechnung statistischer Maßzahlen wie die Summe oder das arithmetische Mittel mehrerer Variablen für einen einzelnen Fall fallen unter die Aggregation (Janssen und Laatz 1994). Die Aggregation hat das Ziel, umfangreiche Daten durch geeignete Verfahren zu verdichten, um wesentliche Information auf einer höheren Abstraktionseben darzustellen (Cleve und Lämmel 2024).

Datenglättung dient dazu, kurzfristige Schwankungen in Daten zu reduzieren und zugrundeliegende Trends sichtbar zu machen. Diese ist eng mit der Rauscherkennung verbunden (Cleve und Lämmel 2024; García et al. 2015). Es bestehen hier also Überlappungen mit der Datenbereinigung. García et al. (2015) betonen, dass die Datenglättung nur dann zuverlässig gelingt, wenn die ursprünglichen Daten ausreichend strukturelle Informationen enthalten. Letztlich besteht die Grundidee darin, jeden numerischen Wert in der vorhandenen Datenmenge durch idealisierte Werte zu ersetzen, die zum Beispiel durch Regression gewonnen werden kann (Cleve und Lämmel 2024). Es existieren verschiedene Glättungsverfahren, wie beispielsweise die Berechnung von gleitenden Durchschnitten oder gleitenden Medianen (Janssen und Laatz 1994). Ein komplexeres Verfahren ist die T4253-Glättung, die als zusammengesetzte Prozedur mehrere Schritte des Median- und Mittelwertbildens mit unterschiedlichen Spannen umfasst (Janssen und Laatz 1994). Bei der Glättung ist die Behandlung fehlender Werte wichtig, um nicht weitere fehlende Werte in der geglätteten Reihe einzuführen (Janssen und Laatz 1994). Verfahren wie das Binning kommen zum Einsatz, bei dem Datenpunkte in Intervalle eingeteilt und durch deren Mittelwerte ersetzt werden (Cleve und Lämmel 2024).

Generalisierung hat zum Ziel, Daten so umzuwandeln oder zusammenzufassen, dass der anschließende Data-Mining-Schritt effizienter durchgeführt werden kann. Dabei kommen häufig Transformationsverfahren zum Einsatz, die eine menschliche Überwachung erfordern und stark von den Eigenschaften der zugrunde liegenden Daten abhängen (García et al. 2015). Die Verfahren der Generalisierung beinhalten nach (Janssen und Laatz 1994) die Zusammenfassung von Kategorien beziehungsweise die Bildung von Klassen bei metrischen Daten durch

Umkodieren. Eine weitere Form stellt die Zusammenfassung mehrerer Werte oder großer Wertebereiche zu Werteklassen dar, die sich ebenfalls durch Umkodieren realisieren lässt (Janssen und Laatz 1994). Die Generalisierung lässt sich als Anpassung auf der Informationsebene beschreiben, weil sie den Detaillierungsgrad der Daten verändert und sie auf einer höheren Abstraktionsebene darstellt (García et al. 2015). Eng verwandt mit der Generalisierung ist die

Generierung von Konzepthierarchien ist eng mit der Generalisierung verwandt. Sie ermöglicht bei nominalen Daten die Generalisierung von Merkmalen wie „Straße“ zu höherstufigen Konzepten wie „Stadt“ oder „Land“ (Han 2012). Solche Hierarchien können manchmal implizit im Datenbankschema vorhanden sein und automatisch definiert werden (Han 2012).

Die Notwendigkeit der Datentransformation lässt sich im Wesentlichen auf zwei Bereiche zurückführen:

1. *Algorithmische Kompatibilität*: Viele Data-Mining-Verfahren sind auf spezifische Datenformate angewiesen (García et al. 2015). Entsprechend bezieht sich die algorithmische Kompatibilität auf die Anforderungen, die statistische Verfahren und Data-Mining-Algorithmen an die Daten stellen, um korrekt und effizient zu funktionieren (Borrohou et al. 2025). Um bestimmte Algorithmen korrekt ausführen zu können, müssen Daten in eine verarbeitbare Form überführt werden. Dies erfordert, dass die algorithmische Kompatibilität auf Datenebene gegeben ist. So wird sichergestellt, dass die Anforderungen des jeweiligen Algorithmus an Format, Datentyp, Wertebereich und Struktur der Eingabedaten erfüllt sind. Zwei grundlegende Annahmen betreffen oft die Normalverteilung der Variablen sowie eine gleichbleibende Varianz über den gesamten Wertebereich (Osborne 2010).

2. *Informationsoptimierung*: Die Informationsoptimierung zielt darauf ab, die Datenqualität zu verbessern, um die Interpretation der Daten zu erleichtern (Manikandan 2010). So kann die Datentransformation die Verständlichkeit und Zugänglichkeit von Informationen verbessern, beispielsweise durch die Umwandlung von Datumsvariablen, um präzisere Zeitangaben oder die Berechnung von Zeitintervallen zu ermöglichen (Janssen und Laatz 1994). Strukturelle Transformationen dienen dazu, den Aufbau eines Datenbestands an die Anforderungen spezifischer Analysen anzupassen (Borrohou et al. 2025). Die Umkodierung von Werten zur Zusammenfassung von Kategorien oder die Erstellung neuer Variablen durch das Zählen bestimmter Merkmalsausprägungen dient dazu, die Daten gezielt für Analysezwecke zu optimieren und neue Perspektiven auf Zusammenhänge zu ermöglichen (Janssen und Laatz 1994). Die Informationsoptimierung findet somit primär auf der Informationsebene statt, indem die Datenqualität, Konsistenz und Nützlichkeit der im Datenbestand repräsentierten Informationen verbessert wird (Cleve und Lämmel 2024).

Die Auswahl der zu vergleichenden Datentransformationsverfahren orientiert sich an dem methodischen Prinzip, dass nur solche Verfahren miteinander verglichen werden, die ein identisches Ziel verfolgen und zu unterschiedlichen, messbaren Ergebnissen führen. An dieser Stelle wird sich für diese Arbeit auf Transformationen, die auf Datenebene ablaufen und somit strukturelle Anpassungen der Daten vornehmen, festgelegt.

3.3 Grundlagen der Datenreduktion

Die wachsende Datenmenge im Zuge der Digitalisierung stellt die Analyse dieser vor komplexen Herausforderungen. Neben dem steigenden Speicherbedarf erschwert insbesondere die hohe Dimensionalität moderner Datensätze die effiziente Extraktion verwertbarer Erkenntnisse. Dieses Phänomen, bekannt als „Fluch der Dimensionalität“, führt dazu, dass die Leistung von Algorithmen mit zunehmender Merkmalsanzahl drastisch sinkt, während der Ressourcenbedarf wächst (Runkler 2010). Gleichzeitig gefährden redundante oder inkonsistente Daten die Validität von Analyseergebnissen. Fehlerhafte Eingaben produzieren zwangsläufig fehlerhafte Ausgaben (Knobloch 2001). Vor diesem Hintergrund etabliert ist das Ziel der Datenreduktion, die Datenmenge zu verringern ohne wesentliche Informationen zu verlieren (Han 2012). Während Methoden wie die Dimensionsreduktion Rechenzeit und Speicherbedarf reduzieren, bergen sie das Risiko, interpretierbare Merkmale in abstrakte Komponenten zu überführen, was die Nachvollziehbarkeit erschwert (Berry und Linoff 2009). Andere Ansätze wie die Instanzenreduktion filtern repräsentative Teilmengen aus großen Datenbeständen, um die Analyse auf relevante Fälle zu fokussieren (Cleve und Lämmel 2024).

3.4 Verfahren der Datenreduktion

Ausgehend von den theoretischen Grundlagen lassen sich die Methoden der Datenreduktion systematisch in zwei Hauptkategorien untergliedern, die Merkmalsreduktion und die Instanzenreduktion (Aggarwal 2014).

Merkmalsreduktion

Die Merkmalsreduktion adressiert hochdimensionale Daten, indem redundante oder irrelevante Attribute identifiziert und entfernt werden. Im Wesentlichen wird Merkmalsreduktion als die Umwandlung einer hochdimensionalen Darstellung von Daten in eine niedrigdimensionale Darstellung beschrieben (van der Maaten et al. 2007). Dabei wird sich auf die Verringerung der Anzahl der Variablen oder Attribute im Datenbestand bezogen (Ayeche und Alti 2023). Ayeche und Alti (2023) und Li et al. (2024) gliedern die Merkmalsreduktion in die Dimensionsreduktion und in die Merkmalsauswahl.

Dimensionsreduktion ist ein zentraler Ansatz, mit der Daten in einen niedrigdimensionalen Raum transformiert werden, wobei möglichst viel Information erhalten bleibt (Jolliffe 2011).

Diese umfasst Techniken, die die ursprünglichen Merkmale in einen neuen, kleineren Satz von Merkmalen umwandeln (Ayeche und Alti 2023). In der Literatur werden Verfahren innerhalb der Dimensionsreduktion weiter eingeteilt in lineare und nicht-lineare Techniken (Ayeche und Alti 2023; Wang et al. 2024). Lineare Verfahren wie die Hauptkomponentenanalyse (PCA) projizieren Daten orthogonal auf unkorrelierte Hauptkomponenten, die maximale Varianz erklären (Li et al. 2024). PCA läuft unüberwacht ab und basiert auf einer Eigenvektorsuche. Diese maximiert die Varianz innerhalb der ausgewählten Komponenten (Hardman und Tapamo 2024). Das Verfahren kann die Gesamtcharakteristik der Daten besser erhalten (Wang et al. 2024). Die Faktoranalyse (FA) wird zur Validierung von Fragebögen eingesetzt, indem sie die Kovarianzstruktur beobachtbarer Variablen durch eine geringere Anzahl Faktoren modelliert, die als gemeinsame Einflussgrößen auf die Variablen wirken (You und Hung 2021; Uhm et al. 2012) . Die unabhängige Komponentenanalyse bildet hochdimensionale Daten in einen Raum unabhängiger Komponenten ab, währen Redundanz und Korrelation der Daten minimiert werden (Wang et al. 2024). Die lineare Diskriminanzanalyse (LDA) ist eine überwachte Lernmethode. Sie bildet hochdimensionale Daten in einen niedriger dimensionalen Raum ab, während die Klassenunterschiede maximiert werden (Wang et al. 2024). Nichtlineare Dimensionsreduktionsverfahren bilden hochdimensionale Daten mithilfe nicht-linearer Transformationen in einen niedriger dimensionalen Raum ab (Wang et al. 2024). Für nichtlineare Datenstrukturen bieten Methoden wie die Kernel-PCA (KPCA) (Hofmann et al. 2008) Lösungen, die lokale Ähnlichkeiten bewahren oder komplexe Muster visualisieren. KPCA ermöglicht PCA die Verarbeitung nicht-linearer Beziehungen, indem Daten in einen höherdimensionalen Raum abgebildet werden (Hardman und Tapamo 2024). Die kategoriale PCA wird als geeigneter für kategoriale Daten oder Ordinaldaten betrachtet als PCA oder FA (Campos et al. 2020). Die Idee besteht darin, für jede Kategorie einen optimalen Zuweisungswert zu ermitteln, um ein bestimmtes Kriterium zu erfüllen (Campos et al. 2020). Das Verfahren t-distributed Stochastic Neighbor Embedding (t-SNE) zeichnet sich durch die Erhaltung lokaler Beziehungen aus und ist leistungsfähig für Visualisierung und Clustering (Villares et al. 2024). Die Independent Component Analysis (ICA) dient dazu vermischte Signale in seine reduzierten Unterkomponenten aufzuteilen (Joshi und Machchhar 2014). ICA generiert neue Merkmale aus den ursprünglichen Daten, anstatt eine Untermenge der bestehenden Merkmale auszuwählen (Zebari et al. 2020). FastICA ist ein gängiges iteratives ICA-Verfahren (Wang et al. 2024). FastICA verwendet als Zielfunktion die Maximierung der Nicht-Gauß'schen Verteilung der Daten, da angenommen wird, dass die gesuchten unabhängigen Komponenten nicht normalverteilt sind (Yang 2024).

Merkmalsauswahl selektiert eine Teilmenge relevanter Merkmale direkt, ohne deren semantische Bedeutung zu verändern (Petersohn 2005). Die Auswahl geeigneter Merkmale kann

dazu beitragen, die Verarbeitungsgeschwindigkeit zu erhöhen und Zeit und Aufwand zu reduzieren (Ayeche und Alti 2023). Die Untermenge der Eingabemerkmale wird ausgewählt. Dies sind die relevantesten und nicht-redundanten Merkmale, ohne dass eine Datenumwandlung vorgenommen wird (Bahri et al. 2020). Während die Dimensionsreduktion bei nicht interpretierbaren Rohdaten bevorzugt wird, ist die Merkmalsauswahl in domänenspezifischen Anwendungen kritisch (Hippner et al. 2011). AlKarawi und AlJanabi (2022) unterscheiden in Filter-, Wrapper- und Embedded-Methoden. Die Wahl des Verfahrens hängt vom Kontext ab. Filtermethoden bewerten Merkmale unabhängig vom Lernalgorithmus, sind schnell, aber ignorieren Modellinteraktionen. Wrapper-Methoden nutzen den Lernalgorithmus hingegen zur Bewertung und liefern oft bessere Ergebnisse, sind jedoch rechenintensiv. Embedded-Methoden integrieren die Merkmalsauswahl ins Modelltraining und sind effizienter und robuster gegenüber Überanpassung (AlKarawi und AlJanabi 2022). Der Chi-Quadrat-Test ist ein statistischer Hypothesentest, der verwendet wird, um festzustellen, ob zwei kategoriale oder nominale Variablen wahrscheinlich verwandt sind oder nicht (Ghoul et al. 2024). Minimum Redundancy Maximum Relevance (MRMR) hat das Ziel, Merkmale auszuwählen, die eine hohe Relevanz für die Zielvariable aufweisen und gleichzeitig minimale Redundanz untereinander besitzen (Ghoul et al. 2024). Campos et al. (2020) führt auch die Pearson-Korrelation als statistisches Maß an. Diese kann genutzt werden um lineare Zusammenhänge zwischen Variablen zu messen. Auch der Autoencoder wird hier als Verfahren genannt und rekonstruiert Daten durch ein komprimiertes Zwischenformat.

Instanzenreduktion

Die Instanzenreduktion wird nach (Aggarwal 2014; García et al. 2015; Han 2012; Cleve und Lämmel 2024) auch als numerische Datenreduktion oder nur Reduktion der Datenmenge genannt. Sie befasst sich im Wesentlichen damit die Anzahl der Beobachtungen oder Datenpunkte zu reduzieren, ohne wesentliche Informationen zu verlieren (Han 2012). Es wird zwischen Sampling-Verfahren und Verfahren zur Instanzenauswahl unterschieden.

Sampling-Verfahren extrahieren repräsentative Teilmengen, wobei wahrscheinlichkeitsbasierte Methoden wie geschichtete Stichproben die Repräsentativität durch gezielte Berücksichtigung von Häufigkeitsverteilungen sichern (Cleve und Lämmel 2024). Jeder Datenwert hat also die gleiche Auswahlwahrscheinlichkeit (Biswas et al. 2022). Die Teilmengen werden so ausgewählt, dass sie den gesamten Datensatz repräsentieren (Biswas et al. 2022; AlKarawi und AlJanabi 2022). Nicht-wahrscheinlichkeitsbasierte Ansätze wie Schneeballverfahren bieten Flexibilität, bergen jedoch Risiken verzerrter Ergebnisse (Petersohn 2005). Nach AlKarawi und AlJanabi (2022) können die wahrscheinlichkeitsbasierten und nicht-wahrscheinlichkeitsbasierten Ansätze in vier Kategorien aufgeteilt werden. Hierzu gehört einfaches Zufalls-Samp-

ling ohne Zurücklegen, einfaches Zufalls-Sampling mit Zurücklegen, Clustersampling und geschichtetes Zufalls-Sampling, bei dem die Daten zunächst in Schichten oder Gruppen unterteilt werden. Biswas et al. (2022) nennt außerdem das datengesteuerte Sampling, das die lokalen Eigenschaften der Daten, wie Skalarwert oder multivariate Assoziation, berücksichtigt. Verfahren hierin sind Pointwise Mutual Information für multivariate Datensätze, um statistische Assoziationen zu erfassen oder Histogramm-basiertes Sampling. Unter- oder Über-Sampling wird eingesetzt, um Klassenungleichgewichten in Datensätzen zu adressieren und Verteilung zu harmonisieren (Li et al. 2024).

Instanzenauswahl identifiziert informative Datensätze, entweder durch statistische Filter (Aggarwal 2014). Clustering gruppiert ähnliche Instanzen und selektiert Repräsentanten wie Clusterzentren, was besonders bei natürlichen Datenstrukturen effektiv ist (Xu und Wunsch 2005).

Vergleichsstudien

Es gibt zahlreiche Vergleichsstudien in der Literatur, die unterschiedliche Verfahren der Merkmalsreduktion miteinander vergleichen (van der Maaten et al. 2007). Hauptsächlich wird die Leistung von Klassifikatoren, wie Entscheidungsbäume, k-NN, SVM oder neuronale Netze, nachdem die Merkmalsreduktion angewendet wurde, verglichen. Hierzu wurde die Klassifikationsgenauigkeit oder die Fehlklassifizierungsrate gemessen. Zum Beispiel wurde die Genauigkeit von SVM und neuronalen Netzen verglichen oder die Genauigkeit verschiedener FS-Algorithmen mit einem SVM-Klassifikator. So betrachtet auch Campos et al. (2020) wie sich die Anwendung verschiedener Verfahren der Datenreduktion auf die Leistung in nachfolgenden Schritten wie Klassifikation oder Clustering auswirkt. Die Leistung nachfolgender Data-Mining-Aufgaben wie Klassifikation und Clustering wird in der Literatur umfassend verglichen (Song et al. 2013; Sellami und Farah 2018; Bartenhagen et al. 2010; N. Sharma und K. Saroha 2015). Oft wird die Leistung auf dem Originaldatensatz als Basislinie herangezogen (Ghoul et al. 2024; Song et al. 2013; Sellami und Farah 2018). Van der Maaten et al. (2007) nutzen auch die Berechnungszeit für den Vergleich. Diese wird direkt gemessen und verglichen, um Aussagen zur Effizienz der Verfahren zu treffen. Es wird also die tatsächliche Zeit, die für die Durchführung der Reduktion benötigt wird gemessen (Ayeche und Alti 2023), um eine theoretische Bewertung der algorithmischen Komplexität zu ermöglichen (AlKarawi und AlJanabi 2022). Die tatsächliche Zeit, die das Reduktionsverfahren oder der gesamte Prozesse, also Reduktion mit anschließendem Modelltest, benötigt, wird gemessen (Sellami und Farah 2018). Alternativ oder ergänzend wird die theoretische algorithmische Komplexität bewertet (Sellami und Farah 2018; van der Maaten et al. 2007). Ein anderer Vergleich erfolgt auch auf Informationsebene. Der Informationserhalt wird geprüft, indem der Rekonstruktionsfehler gemessen wird (Wang et al. 2024). Es wird hierin versucht die Originaldaten aus den reduzierten Daten zu rekonstruieren und die Differenz zu berechnen. Bei PCA wird der Anteil der Gesamtvarianz

berechnet, der durch die ausgewählten Hauptkomponenten erhalten bleibt (Wang et al. 2024). Villares et al. (2024) misst wie gut die ursprünglichen Deskriptoren anhand der reduzierten Daten vorhergesagt werden können. Eine spezifische Metrik ist beispielsweise der Corrected Rand Index, der gefundene Cluster mit einem Referenz-Cluster vergleicht. Ein weiterer Aspekt für den Vergleich ist der Erhalt der Datenstruktur, bei der geprüft wird wie gut die Beziehungen zwischen den Datenpunkten, sowohl lokal als auch global, im reduzierten Raum erhalten bleiben. Hierzu wird Pearson-Korrelation oder Clustering-Ähnlichkeit genutzt, um die Struktur in den Originaldaten mit denen in den reduzierten Daten zu vergleichen (Wang et al. 2024). Wang et al. (2024) und Villares et al. (2024) untersuchen auch die Interpretierbarkeit und wie die reduzierten Dimensionen oder Komponenten verstanden werden können. Die qualitative Bewertung erfolgt durch Experten des jeweiligen Fachgebietes. In der Literatur werden außerdem auch die Visualisierungsqualität, durch Inspektion der generierten Diagramme, Robustheit gegenüber Rauschen (Ayeche und Alti 2023) und die Eignung für bestimmte Datentypen oder Verteilungen untersucht (AlKarawi und AlJanabi 2022). Dieser Vergleich der inhärenten Eigenschaften basierend auf Datentypeignung, Skalierbarkeit, Robustheit gegenüber Rauschen, Rechenkomplexität, ob Merkmale parametrisch sind und welche Art von Objektivfunktion sie optimieren, wird oft verglichen (Bartenhagen et al. 2010; van der Maaten et al. 2007; S. K. Joshi und S. Machchhar 2014). Die Vergleiche erfolgen basierend auf den Eigenschaften wie Linearität, Datenverteilungsannahmen, Datentypeignung, Parametrizität, Lerntyp und ob diese deterministisch sind (Ayeche und Alti 2023). Hierbei wird auch die Skalierbarkeit und der Umgang mit Rauschen untersucht.

Im Rahmen der Instanzenreduktion untersucht (Biswas et al. 2022) die Fähigkeit der Sampling-Verfahren, die Merkmale der Daten und die statistischen Assoziationen beziehungsweise Korrelationen zwischen Variablen in multivariaten Datensätzen zu erhalten. Hierin wird für die Qualität der Stichproben der Pearson-Korrelationskoeffizient verwendet, um visualisierte Bilder aus gesampelten Daten mit Bildern aus den Originaldaten zu vergleichen. Dies geschieht auf Basis der RGB-Kanäle der gerenderten Bilder (Biswas et al. 2022). In anderen Studien wird der Einfluss auf Ergebnisvariablen wie Tragedauer oder durchschnittliche Aktivität untersucht (Mâsse et al. 2005). Hierfür werden die Mittelwerte und Standardabweichungen dieser Variablen analysiert, die mithilfe unterschiedlicher Berechnungsmethoden ermittelt wurden. Auch werden Metriken wie die Ein-/Ausgabezeit (Mâsse et al. 2005) oder die Ausführungszeit gemessen und verglichen (Ratnoo 2013). Die Leistung wird ebenfalls über einen Klassifikator auf dem reduzierten Datensatz getestet, wobei Ratnoo (2013) nicht explizit benennt welche verschiedenen Klassifikatoren angewandt werden. Als Vergleichsmetriken werden die resultierende Genauigkeit und der mittlere absolute Fehler des Klassifikators verwendet (Ratnoo 2013). Auch hier wird wieder mit den Modellen verglichen, bei der der Datenbestand ohne

Instanzenreduktion genutzt wird. AlKarawi und AlJanabi (2022) bewerten die allgemeine Eignung von Verfahren anhand Datentyps, Datensatzgröße, Anwendungsziel, Vorhandensein von Rauschen oder Ausreißern und dem Kompromiss zwischen reduzierten Daten und benötigtem Wissen. Hierbei handelt es sich jedoch um konzeptionelle Bewertungen und nicht um eine quantifizierbare Vergleichskennzahl.

Die verschiedenen Aufgaben der Datenvorverarbeitung, Datenintegration, Datenbereinigung, Datentransformation und Datenreduktion, sind direkte Antworten auf die Diskrepanz zwischen Rohdaten und Algorithmusanforderungen (Michael Goebel und Le Gruenwald 1999). Darüber hinaus erfordert die Datenvorverarbeitung eine intensive Unterstützung durch den Analysten (Cleve und Lämmel 2024). Entscheidungen darüber, welche Techniken angewendet werden sollen, basieren nicht nur auf den zugrundeliegenden Daten und den Eingabeanforderungen des ausgewählten Algorithmus, sondern erfordern auch Wissen über die anzuwendenden Verfahren und das Fachgebiet (Han 2012).

4 Entwicklung eines experimentellen Vergleichrahmens

Nachdem Kapitel 2 und 3 die theoretischen Grundlagen der Wissensentdeckung in Datenbanken, des Data Mining sowie die Aufgaben und Verfahren der Datentransformation und Datenreduktion dargelegt haben, konzentriert sich das vorliegende Kapitel auf die Entwicklung eines experimentellen Vergleichrahmens. Dessen Ziel ist es, einen systematischen Vergleich der Verfahren der Datentransformation und Datenreduktion zu ermöglichen. Hierzu wird ein methodisches Vorgehen verfolgt, das von der Ableitung eines Kriterienkatalogs und der Entwicklung geeigneter Metriken bis zur Entwicklung des experimentellen Vergleichrahmens führt.

4.1 Methodisches Vorgehen

Um das Ziel der Entwicklung eines experimentellen Vergleichrahmens strukturiert zu erreichen, wird im Folgenden zunächst das methodische Vorgehen erläutert. Der experimentelle Vergleichsrahmen ermöglicht einen systematischen Vergleich der Verfahren der Datentransformation und Datenreduktion. Die Entwicklung dessen erfolgt in einem mehrstufigen Prozess, der auf einer kritischen Synthese und Operationalisierung von Randbedingungen und Anforderungen des Data Mining und der Datenvorverarbeitungsverfahren selbst, sowie auf bestehender Literatur aufbaut. Schritt für Schritt werden daraus die erforderlichen Bausteine entwickelt, die schlussendlich in die Konzeption des Vergleichrahmens einfließen.

Für das methodische Fundament wird ein Kriterienkatalog abgeleitet. Dieser soll eine strukturierte Herangehensweise gewährleisten und die darauf basierenden Metriken und den Vergleichsrahmen selbst methodisch absichern. Anschließend kommt es zu der Operationalisierung der Anforderungen, sodass die Entwicklung der Metriken die Anforderungen der nachgelagerten Data-Mining-Algorithmen berücksichtigt. Für die Festlegung geeigneter Rahmenbedingungen sind insbesondere die Eigenschaften der Rohdaten von Bedeutung. Auch diese tragen dazu bei, dass im Folgeschritt die messbaren Metriken entwickelt werden. Diese dienen zum einen der Bewertung von Datentransformations- und Datenreduktionsverfahren in isolierten Anwendungsszenarien. Zum anderen werden sie im Hinblick auf den Reihenfolgeeffekt auch für die kombinierte Anwendung beider Verfahren entwickelt, bei der Datentransformation und Datenreduktion nacheinander ausgeführt werden. Durch diese duale Betrachtungsweise können sowohl die individuellen Eigenschaften der Vorverarbeitungsverfahren als auch deren Interaktionseffekte bei sequenzieller Anwendung erfasst werden. Um schließlich sowohl die isolierte als auch kombinierte Bewertung der Datentransformations- und Datenreduktionsverfahren umsetzen zu können, wird in einem letzten Schritt der experimentelle Vergleichsrahmen zusammengestellt.

4.2 Ableitung eines Kriterienkatalogs

Die Entwicklung des experimentelle Vergleichsrahmens wird anhand eines Kriterienkatalogs umgesetzt. Zum einen stellen die Kriterien sicher, dass sowohl die grundlegende Anwendbarkeit als auch die differenzierte Leistungsbewertung der Verfahren angemessen berücksichtigt werden. Zum anderen definieren die Kriterien klar nach welchen Maßstäben die Metriken für die Datentransformations- und Datenreduktionsverfahren bewertet werden. Die Kriterien gliedern sich in drei komplementäre Ebenen, nämlich Einhaltungs-, Evaluations- und Einflusskriterien. Einhaltungskriterien legen die technischen Voraussetzungen fest, unter denen ein Verfahren angewendet werden kann. Sie wirken somit wie ein binärer Filter, der Verfahren entweder als anwendbar oder nicht anwendbar klassifiziert. Ist diese Grundlage erfüllt, ermöglichen Evaluationskriterien die quantitativen Bewertungen der Effektivität eines Vorverarbeitungsverfahrens. Evaluationskriterien repräsentieren messbare Zielgrößen und sind endergebnisorientiert, da sie den direkten Erfolg der Datenvorverarbeitung im Kontext von Data Mining quantifizieren. Einflusskriterien hingegen beschreiben die Randbedingungen, die die Wirksamkeit der Vorverarbeitungsverfahren begünstigen oder einschränken können. Während Einhaltungskriterien eine harte, binäre Schwelle darstellen, beschreiben Einflusskriterien eine Modulation der Wirksamkeit. In diesem Kapitel werden die unterschiedlichen Kategorien von Kriterien beschrieben sowie eine strukturierte Übersicht erstellt, die als konzeptionelle Grundlage für die anschließenden Weiterentwicklung dient.

Einhaltungskriterien

Die Charakteristika der Rohdaten selbst stellen eine grundlegende Herausforderung für Verfahren der Datenvorverarbeitung dar. Besonders problematisch sind enorme Datenmengen sowie eine unvollständige Form und unstrukturierte oder inkompatible Formate. Diese initialen Eigenschaften der Rohdaten bilden die erste Ebene, die die Anwendbarkeit bestimmter Verfahren beeinflusst. Einhaltungskriterien leiten sich primär aus den Spezifikationen der Vorverarbeitungsverfahren, den mathematischen Anforderungen der zugrundeliegenden Algorithmen sowie technischen Implementierungsrestriktionen ab. Liegen Rohdaten in ein für die Analysezwecke ungeeignetes Format vor, so muss eine Datentransformation erfolgen. Solche Transformationen sind jedoch nicht immer verlustfrei möglich, insbesondere wenn keine natürliche Ordnung zwischen den Ausprägungen besteht (vgl. Abschnitt 3.1). Kategoriale Merkmale, wie etwa Farben, können mithilfe geeigneter Kodierungsverfahren in numerische Formate überführt werden, jedoch gilt dies nicht uneingeschränkt für alle Arten von Rohdaten. Beispielsweise lassen sich Fehlermeldungen in Textform ohne vorherige inhaltliche Strukturierung oder semantische Analyse nur schwierig in ein für Data-Mining-Algorithmen verwertbares Format transformieren. Insofern ist nicht jede Formatumwandlung trivial oder verlustfrei

möglich. Die Untersuchung der Einhaltungskriterien bewegt sich, anders als bei Evaluationskriterien, im Rahmen der Randbedingungen und Voraussetzungen für die Verfahren. Die Einhaltungskriterien sind zwar essenziell, um zu entscheiden, ob und wie ein Datenvorverarbeitungsverfahren überhaupt angewendet werden kann, sie lassen sich jedoch in der Regel nicht direkt als quantitatives Kriterium nutzen. Die grundsätzliche Durchführbarkeit von Verfahren kann mithilfe dieser Einhaltungskriterien geprüft werden. Sie sind notwendig, um zu klären, welche Verfahren für spezifische Datensätze in Frage kommen. Erst wenn technisch umsetzbare Ansätze berücksichtigt werden, kann sichergestellt werden, dass der Vergleichsrahmen aussagekräftige Ergebnisse liefert.

Evaluationskriterien

Evaluationskriterien werden in erster Linie aus den Anforderungen des Data Mining abgeleitet und sind endergebnisorientiert. Sie repräsentieren quantitative Zielgrößen, anhand derer der Erfolg eines Datenvorverarbeitungsschrittes gemessen werden kann. Solche quantitativen Marker umfassen etwa die Modellgenauigkeit eines nachgelagerten Data-Mining-Algorithmus nach Anwendung der Vorverarbeitung oder die benötigte Rechenzeit für den nachfolgenden Analyseprozess. Die Relevanz der Evaluationskriterien liegt darin, dass die Datenvorverarbeitung als erfolgreich angesehen wird, wenn sie entweder den Rohdatenbestand für den Data-Mining-Algorithmus überhaupt erst anwendbar macht oder dessen Leistung verbessert. Die spezifische Auswahl der Evaluationskriterien muss dabei berücksichtigen, welche Aspekte der Datenqualität oder Datenstruktur für den spezifischen Data-Mining-Algorithmus von Bedeutung sind. Hinweise für sinnvolle Evaluationskriterien können hierbei aus Aussagen der bestehenden Literatur gewonnen werden, die aus Abschnitt 2.4 abgeleitet werden. Die Anwendung von Data-Mining-Algorithmen setzt voraus, dass die Eingabedaten spezifische Anforderungen an Struktur und Format erfüllen. Wenn diese Kriterien nicht erfüllt sind, resultieren daraus operationale Schwierigkeiten. Die Datenvorverarbeitung dient gerade dazu, diese Diskrepanz zwischen Rohdaten und Algorithmusanforderungen zu überwinden (vgl. Abschnitt 2.5). Für den experimentellen Vergleichsrahmen werden Data-Mining-Algorithmen betrachtet, die zum einen den Datenvorverarbeitungsschritt der Datentransformation benötigen, als auch diejenigen, die von der Datenreduktion profitieren. Nachfolgend werden die Anforderungen, die an die vorverarbeiteten Daten gestellt werden und unmittelbar zu Evaluationskriterien führen, übersichtlich dargelegt. Sie können den Erfolg der Vorverarbeitung im Kontext von Data Mining quantifizieren.

Während Verfahren wie ID3, C4.5 oder Regel-Lernverfahren nominale oder diskretisierte Daten erfordern, benötigen umgekehrt Algorithmen wie k-NN oder neuronale Netze numerische oder metrische Daten (vgl. Abschnitt 2.4). Rohdaten, die nicht im benötigten Format vorliegen,

erfordern eine entsprechende Transformation wie beispielsweise Diskretisierung oder Binärisation (vgl. Abschnitt 3.2). Verfahren, die auf Distanzmetriken basieren oder Annahmen über die Datenverteilung treffen, reagieren besonders empfindlich auf unterschiedliche Wertebereiche und Verteilungen der Merkmale. Eine ungeeignete Skalierung der Rohdaten für das Zielverfahren kann die Analyse verzerren. Transformationen wie Normalisierung oder Standardisierung passen die Wertebereiche an, um die Vergleichbarkeit der Merkmale zu gewährleisten. Die Datenreduktion kann die Datenmenge verringern und die Komplexität reduzieren, während wichtige Informationen erhalten bleiben (vgl. Abschnitt 3.3).

Einflusskriterien

Einflusskriterien erfassen Kontextfaktoren, die auf die Wirksamkeit von Verfahren einwirken, ohne selbst Gegenstand der Bewertung zu sein. Beispielsweise erschwert eine hohe Dimensionalität die Normalisierung. Gegebenenfalls verlieren dann statistische Kennwerte wie Mittelwert und Standardabweichung ihre Eindeutigkeit und büßen an Zuverlässigkeit ein. Einflusskriterien sind relevant für die Charakteristika der Rohdaten und beschreiben die Ausgangsbedingungen, wie etwa starkes Rauschen im Datensatz, das beispielsweise die Effektivität der Datenglättung beeinflussen kann (vgl. Abschnitt 3.2). Diese Kriterien sind nicht endresultatorientiert und messen nicht die Leistungsfähigkeit eines Verfahrens, sondern unter welchen Bedingungen es funktioniert. Somit sind sie also nicht direkt vergleichbar und lassen sich nicht in dieselben Einheiten übertragen wie es gewisse Leistungskennzahlen können. Sie basieren auf theoretischen Überlegungen zu Anwendbarkeitsgrenzen sowie auf Erkenntnissen aus Vergleichsstudien (vgl. Abschnitt 3.4). Zu den Einflusskriterien zählen auch Faktoren, die sich aus Aspekten der Datenqualität ergeben. An dieser Stelle zeigt sich jedoch die Schwierigkeit, Einflusskriterien eindeutig von Einhaltungskriterien abzugrenzen. Während letztere eine klare Grenze darstellen, beschreiben Einflusskriterien eine kontinuierliche Anpassung der Wirksamkeit. Sie geben an, unter welchen Voraussetzungen ein Verfahren effizient sein kann, ohne dass ihre Verletzung zwangsläufig zur Unbrauchbarkeit des Verfahrens führt. Eine Übersicht zu den Arten von Kriterien und ihren Definitionen sind in der Tabelle 1 zu sehen.

Der in Tabelle 1 zu sehender Kriterienkatalog, stellt die strukturierten Bewertungsmaßstäbe für die drei Kategorien Einhaltungs-, Evaluations- und Einflusskriterien bereit. Jede Kategorie leistet einen spezifischen Beitrag zur Entwicklung des Vergleichsrahmens und der entsprechenden Metriken. Einhaltungskriterien definieren technische Mindestanforderungen, Evaluationskriterien messen die Verfahrensleistung quantitativ und die Einflusskriterien erfassen relevante Rahmenbedingungen. Der Kriterienkatalog dient als Grundlage für die Operationalisierung messbarer Metriken und strukturieren den experimentellen Vergleichsrahmen.

Tabelle 1: Kriterienkatalog

Kriterienart	Definition	Beinhaltet	Charakteristik
Einhaltungskriterien	Mindestvoraussetzung für technische Anwendbarkeit	Kompatibilität bzgl. Datentyp, strukturelle Verfügbarkeit, Kompatibilität bzgl. Format	Erfüllt/Nicht erfüllt
Evaluationskriterien	Quantitative Leistungsmaße zur Bewertung des Verfahrenserfolg	Güte der Vorverarbeitung, Recheneffizienz, nachgelagerte Modellleistung, Informationserhaltung	Messbare Zielgröße
Einflusskriterien	Kontextuelle Faktoren, die Verfahrenswirksamkeit beeinflussen	Datenqualität, Dimensionalität, Domänenspezifität	Graduelle Wirkung

Der vorgestellte Kriterienkatalog strukturiert die qualitativen Bewertungsdimensionen. Um nun objektive und messbare Vergleichbarkeit zu gewährleisten, bedarf es in einem nächsten Schritt einer präzisen Quantifizierung. Die Entwicklung der Metriken wird mit Hilfe des Kriterienkatalogs in Kapitel 4.3 abgeleitet.

4.3 Entwicklung von Metriken zur Bewertung von Datentransformations- und Datenreduktionsverfahren

Aufbauend auf dem in Kapitel 4.2 abgeleiteten Kriterienkatalog, fokussiert sich das vorliegende Kapitel auf die Entwicklung spezifischer Metriken zur Bewertung von Datentransformation- und Datenreduktionsverfahren. Diese Metriken sind essentiell für die Durchführung eines systematischen und quantitativen Vergleichs der Datenvorverarbeitungsverfahren hinsichtlich des experimentellen Vergleichrahmens. Diese Auswertungsmetriken sind darauf ausgelegt, die zuvor abgeleiteten Kriterien und Bewertungsmaßstäbe messbar zu machen, um so einen Vergleich zu ermöglichen.

Operationalisierung der Anforderungen

In dem folgenden Kapitelabschnitt wird der Kriterienkatalog aus 4.2 mehrfach referenziert. Für eine systematische Entwicklung von Metriken und zur adäquaten Erfassung der relevanten

Aspekte der Datentransformations- und Datenreduktionsverfahren im Kontext des Data Mining, ist eine Aufarbeitung der Anforderungen, die sich aus dem Stand der Technik ausarbeiten lassen, erforderlich. Eine unmittelbare Ableitung von Metriken aus den einzelnen Anforderungen würde zu einem unvollständigen Vergleichsrahmen führen. Aus diesem Grund ist die Operationalisierung der Anforderungen mithilfe des Kriterienkatalogs aus Kapitel 4.2 erforderlich, um die Bewertungsdimensionen einordnen zu können. Diese Notwendigkeit der Operationalisierung ergibt sich daraus, dass die Anwendung und Wirksamkeit der Datenvorverarbeitungsverfahren maßgeblich von den Eigenschaften der Rohdaten und den spezifischen Anforderungen der nachfolgenden Data-Mining-Algorithmen abhängen. Erst die Zusammenführung legt fest, welche Veränderungen die Metriken messen müssen und wie sie mit den Zielen der Datenvorverarbeitung im Kontext von Data Mining korrespondieren. Auf diese Art und Weise kann somit die Lücke zwischen den konzeptionellen Kriterien und der operativen Entwicklung messbarer Indikatoren überbrückt werden. In Tabelle 2 werden die Anforderungen der Datenvorverarbeitungsverfahren aus der Datenreduktion an die Rohdaten gezeigt.

Tabelle 2 listet die verschiedenen Aufgaben der Datenreduktion auf. In der mittleren Spalte sind entsprechende Beispielverfahren zu jeder Aufgabe genannt. Die rechte Spalte zeigt die jeweiligen Anforderungen an die Rohdaten. Die Einhaltungskriterien der strukturellen Verfügbarkeit und Datentypkompatibilität spiegeln sich in den Anforderungen wider, die Dimensionsreduktionsverfahren an die Merkmalsbeschaffenheit der Rohdaten stellen. Ohne ausreichende Merkmalsanzahl würde der Fluch der Dimensionalität als grundlegende Problemstellung entfallen und die Dimensionsreduktion somit ihre konzeptionelle Berechtigung verlieren (vgl. Abschnitt 3.4). Das Einhaltungskriterium der Datentypkompatibilität führt zur Anforderung numerischer, skalierter Daten. Diese ergibt sich aus den mathematisch-statistischen Grundannahmen der Algorithmen, die auf linearen Algebraoperationen basieren wie PCA, LDA und t-SNE (vgl. Abschnitt 3.4). Kategoriale und ordinale Daten stellen eine Herausforderung dar, die durch das Einhaltungskriterium der Formatkompatibilität adressiert werden. Diese Datentypen sind nicht unmittelbar mit numerischen Dimensionsreduktionsverfahren kompatibel. Die hohe Dimensionalität, die bereits als Einflusskriterium eingeordnet wurde, ergibt sich weiter als Anforderung aus der Problemdefinition selbst. Dimensionsreduktionsverfahren entfalten ihre praktische Relevanz hinsichtlich Effizienzsteigerung von Data-Mining-Algorithmen bei einer ausreichenden Merkmalsdichte der Rohdaten. Spezifische Anforderungen wie die Existenz linearer Korrelationen bei der PCA ergeben sich aus einer Kombination von Einhaltungs- und Einflusskriterien.

Während PCA technisch auch ohne lineare Zusammenhänge anwendbar bleibt, sinkt die erklärte Varianz pro Hauptkomponente signifikant bei fehlenden linearen Zusammenhängen, wodurch das Verfahren ineffektiv wird (vgl. Abschnitt 3.4).

Tabelle 2: Anforderungen der Datenreduktion

Aufgabe	Beispielverfahren	Anforderungen an Rohdaten
Dimensionsreduktion	PCA, KPCA, FA, ICA, LDA, t-SNE	• Merkmale vorhanden • Skalierte, numerische Daten • Kategoriale, ordinale Daten • Hohe Dimensionalität • PCA: Lineare Korrelationen vorhanden
Merkmalsauswahl	Filterverfahren, Wrapper-Verfahren, MRMR	• Attribute vorhanden • Gelabelte Daten • Keine starke Multikollinearität • Kategoriale, nominale Daten
Sampling	Einfaches Zufalls-Sampling, Datengesteuertes Sampling, Unter-/Oversampling	• Unausgeglichene Klassen • Rohdaten müssen Instanzen/Beobachtungen enthalten • Große Anzahl Instanzen
Instanzenauswahl	Prototyping, statistischer Fehler	• Rohdaten müssen Instanzen/Beobachtungen enthalten • Große Anzahl Instanzen

Die Anforderungen für Verfahren der Merkmalsauswahl ergeben sich ebenfalls aus den Einhaltungskriterien der strukturellen Verfügbarkeit und Datentypkompatibilität. Die hierzu folgenden Erkenntnisse beziehen sich auf Kapitel 3.4. Definitionsgemäß wird bei der Merkmalsauswahl eine Teilmenge vorhandener Merkmale identifiziert, die Anforderung einer hinreichenden Menge an Attributen folgt direkt aus dem Einhaltungskriterium der strukturellen Verfügbarkeit. Das Einhaltungskriterium der Kompatibilität bezüglich Formats führt zu der Anforderung nach gelabelten Daten. Wrapper-Verfahren evaluieren Merkmalskombinationen iterativ anhand der Vorhersageleistung überwachter Lernalgorithmen, während Embedded-Verfahren die Merkmalsauswahl direkt in den Lernprozess integrieren. Die Abhängigkeit zu gelabelten Daten lässt sich somit direkt aus der Arbeitsweise dieser Verfahren herleiten. Die Anforderung zur Vermeidung von Multikollinearität leitet sich aus dem Einflusskriterium der Datenqualität ab. Der Einfluss von stark korrelierten Merkmalen lässt sich nicht mehr klar trennen. Statistisch kommt

es zu Überlagerungen, wodurch Signifikanztests verfälscht werden (vgl. Abschnitt 3.4). Zwar entfernt die Merkmalsauswahl redundante Variablen, aber sie kann nur wirken, wenn die Rohdaten nicht von vornherein starke Abhängigkeiten aufweisen. Spezifische Anforderungen hinsichtlich Datentyps ergeben sich aus den Einhaltungskriterien. So setzen korrelationsbasierte Filter kontinuierliche, numerische Daten voraus, während der Chi-Quadrat-Test auf nominale oder diskretisierte Daten beschränkt ist (vgl. Abschnitt 3.4).

Für Verfahren der Aufgaben Sampling und Instanzenauswahl ergeben sich die Anforderungen in erster Linie aus dem Einhaltungskriterium der strukturellen Verfügbarkeit. Die grundlegende Voraussetzung ist das Vorhandensein einer hinreichend großen Anzahl von Instanzen, sodass eine sinnvolle Reduktion ohne kritischen Informationsverlust ermöglicht wird. Ohne ausreichende Instanzenanzahl ist folglich keine repräsentative Reduktion möglich, wodurch sich diese Anforderung direkt aus der Definition der strukturellen Verfügbarkeit ableitet. Für Sampling-Verfahren resultiert mit dem Einflusskriterium der Datenverteilung die Anforderung nach unausgeglichenen Klassen (vgl. Abschnitt 3.4). Die Instanzenauswahl zielt darauf ab eine repräsentative Teilmenge der Rohdaten zu finden, die die ursprüngliche Verteilung möglichst gut approximiert. Hierfür müssen die Rohdaten ausreichend dicht und strukturiert sein, um auch Randbereiche valide abzubilden. Diese Anforderung ergibt sich aus der Kombination des Einhaltungskriteriums der strukturellen Verfügbarkeit mit dem Einflusskriterium der Datenverteilung. Nachdem die Anforderungen der Datenreduktion erläutert wurden, ist in Abbildung 5 dessen Operationalisierung dargestellt.

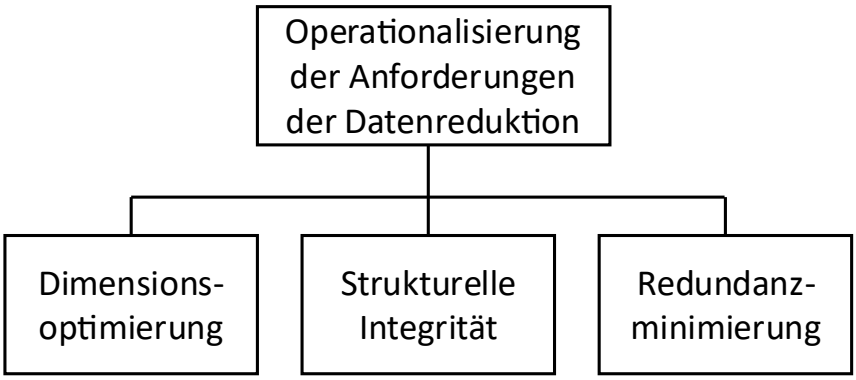

Abbildung 5: Operationalisierung der Anforderungen der Datenreduktion

In Abbildung 5 sind die drei Dimensionen Dimensionsoptimierung, strukturelle Integrität und Redundanzminimierung der Operationalisierung der Anforderungen der Datenreduktion zugeordnet. Die Dimensionsoptimierung ergibt sich aus der expliziten Anforderung einer hohen Dimensionalität bei Verfahren der Dimensionsreduktion ab. Hiermit wird verdeutlicht, dass diese Verfahren auf die effiziente Reduktion hochdimensionaler Datenräume abzielen. Maßgeblich für den Erfolg ist die Fähigkeit, Informationsverlust trotz reduzierter Dimensionalität zu minimieren. Die Varianzerhaltung in den ersten k Hauptkomponenten VE_k kann als Metrik zur

quantitativen Bewertung der strukturellen Stabilität von Daten nach der Vorverarbeitung genutzt werden (vgl. Anhang A). Sie misst, wie viel der ursprünglichen in den ersten k Hauptkomponenten enthaltenen Varianz durch die Datenvorverarbeitung erhalten bleibt. Formal ergibt sich die Metrik als Verhältnis der kumulierten erklärten Varianz der ersten k Hauptkomponenten nach der Vorverarbeitung $\mathrm{Var}_i^{\mathrm{nach}}$ zur kumulierten erklärten Varianz derselben Komponenten vor der Vorverarbeitung $\mathrm{Var}_j^{\mathrm{vor}}$:

$$VE_k = \frac{\sum_{i=1}^{k} \mathrm{Var}_i^{\mathrm{nach}}}{\sum_{j=1}^{k} \mathrm{Var}_j^{\mathrm{vor}}} \tag{1}$$

Die Varianzerhaltung zielt darauf ab, die Bewahrung dominanter Datenstrukturen zu erfassen. Die Berechnungsgrundlage dieser Metrik liegt in der PCA, die die Daten in orthogonale Richtungen projiziert, die nach absteigender Varianz geordnet sind (vgl. Abschnitt 3.4). Wird beispielsweise durch ein Datenvorverarbeitungsverfahren die erklärbare Varianz der wichtigsten k Komponenten von ursprünglich $80\ \%$ auf $72\ \%$ reduziert, so ergibt sich ein VE_k von $0{,}9$. Dies deutet auf eine 90-prozentige Erhaltung der dominanten Strukturen hin. Ein Wert von 1 würde eine vollständige Erhaltung bedeuten und Werte über 1 sind theoretisch möglich würden jedoch auf eine überangepasste oder verzerrte Vorverarbeitung deuten. Die Dimensionsoptimierung kann weiter noch mit der Informationserhaltung IE ausgedrückt werden. Diese basiert auf dem Verhältnis der Mutual Information I (vgl. Anhang A) und bewertet die Erhaltung der Beziehung zwischen Merkmalen und Zielvariablen:

$$IE = \frac{I_{nach}}{I_{vor}} \tag{2}$$

Diese Metrik ist besonders relevant für überwachte Reduktionsverfahren, da sie nichtlineare Abhängigkeiten erfasst. Werte nahe 1 deutet darauf hin, dass die durch die Merkmale getragene Information über das Ziel weitestgehend erhalten geblieben ist. Auch hier können Werte größer als 1 auftreten, wenn die Zielrelevanz der Merkmale durch die Datenvorverarbeitung erhöht wurde. Kritisch hingegen sind Werte im Bereich unter $0{,}5$. Hier liegt erheblicher Informationsverlust vor, der sich wiederum negativ auf die spätere Leistung der Data-Mining-Algorithmen auswirken kann.

Eine weitere Dimension ist die strukturelle Integrität. Anforderungen wie beispielsweise das Vorhandensein linearer Korrelationen für PCA und keiner starken Multikollinearität für Merkmalsauswahlverfahren zeigen, dass Reduktionsverfahren spezifische Struktureigenschaften der Daten voraussetzen und diese bewahren müssen (vgl. Tabelle 2). Die strukturelle Integrität stellt sicher, dass wesentliche Datenbeziehungen trotz der Reduktion erkennbar und nutzbar bleiben. Hierzu misst die Korrelationserhaltung KE wie stark die durchschnittliche Korrelationsstruktur eines Datensatzes durch die Datenvorverarbeitung verändert wird. Die Metrik

basiert auf der Spearman-Rangkorrelation, die auch monotone, nicht-lineare Zusammenhänge erfassen kann (vgl. Anhang A). Formal definiert sich KE als:

$$KE = 1 - |\rho_{vor} - \rho_{nach}| \quad (3)$$

Die absolute Differenz der durchschnittlichen Korrelationen vor und nach der Datenvorverarbeitung quantifiziert die strukturelle Veränderung der Merkmalsbeziehungen. Durch Subtraktion dieses Wertes von 1 ergibt sich eine Metrik, bei der ein Wert von 1 für eine vollständige Erhaltung der ursprünglichen Korrelationsstruktur steht, während ein Wert von 0 auf deren vollständigen Verlust hinweist. Ein Wert von beispielsweise 0,9 bedeutet, dass im Mittel 90 % der ursprünglichen Korrelationen erhalten geblieben sind. Die Dimension zur strukturellen Integrität kann weiter noch mit Hilfe des Silhouetten-Scores bewertet werden (vgl. Anhang A). Diese Metrik bewertet die Erhaltung natürlicher Datengruppierungen nach der Datenreduktion. Der Silhouetten-Score wird in die Bewertung einbezogen, da die Erhaltung der Trennschärfe zwischen verschiedenen Datengruppen insbesondere für Data-Mining-Aufgaben wie die Clusteranalyse von Bedeutung ist (vgl. Abschnitt 2.4). In diesem Zusammenhang wird die Metrik, die gerade den Vergleich des Silhouetten-Scores vor und nach der Datenvorverarbeitung darstellt, ab jetzt als Cluster-Trennbarkeit bezeichnet.

Die letzte Dimension hier ist die Redundanzminimierung. Anforderungen wie gelabelte Daten für Merkmalsauswahlverfahren und große Anzahl Instanzen für Sampling-Verfahren verdeutlichen, dass Reduktionsverfahren auf die gezielte Entfernung überflüssiger Informationen abzielen. Der Erfolg liegt in der Identifikation und Entfernung redundanter Elemente bei gleichzeitiger Bewahrung wichtiger Informationen. Zu dieser Dimension wird die Merkmalsinterdependenz MI wie folgt definiert:

$$MI = \mu(|\rho_{ij}|) \text{ , für } i \neq j \quad (4)$$

Diese Metrik wird als Kennzahl zur Bewertung der durchschnittlichen linearen Abhängigkeit zwischen allen Merkmalspaaren eines Datensatzes nach einer Transformation genutzt. Formal wird sie definiert als Mittelwert der absoluten Korrelationskoeffizienten zwischen allen eindeutigen Merkmalskombinationen, wobei ausschließlich Paare $i \neq j$ berücksichtigt werden. Die Berechnung basiert auf der oberen Dreiecksmatrix der Korrelationsmatrix, um doppelte Zählungen zu vermeiden (vgl. Anhang A). Die Korrelation nach Pearson (vgl. Abschnitt 3.4) zwischen den Merkmalen i und j wird durch ρ_{ij} beschrieben und über den Mittelwert μ über alle betrachteten Paare dargestellt. Ein MI-Wert nahe Null weist auf eine weitgehende Unabhängigkeit der Merkmale hin und signalisiert damit eine erfolgreiche Entkopplung ursprünglich korrelierter Strukturen. Solche Ergebnisse sind insbesondere für Algorithmen von Vorteil, die unter Multikollinearität leiden, wie etwa lineare Regressionsmodelle (vgl. Abschnitt 2.4). Die MI ist eine skaleninvariante Größe, da Korrelationen im normierten Intervall $[-1,1]$ liegen.

Durch die Verwendung des Betrags kann eine richtungsunabhängige Bewertung vorgenommen werden.

Zur Untersuchung der Anforderungen der Datentransformation an die Rohdaten werden zunächst die Aufgaben der Datentransformation betrachtet, die auf Datenebene stattfinden. Die Anforderungen sind in Tabelle 3 dargestellt.

Tabelle 3: Anforderungen der Datentransformation

Aufgabe	Beispielverfahren	Anforderungen an Rohdaten
Normalisierung, Skalierung, Standardisierung	Min-Max-Normalisierung, Z-Wert-Normalisierung, Dezimalskalierung	• Numerische Daten • Kategoriale, ordinale Daten • Keine extremen Ausreißer
Verteilungstransformationen	Logarithmische Transformation, Exponentialtransformation, Box-Cox-Transformation	• Numerische (kontinuierliche bei Box-Cox) Daten • Keine extremen Ausreißer • Positive Werte (bei Log- und Box-Cox-Transformation) • Merkmale mit unterschiedlichen Skalen • Schiefe Verteilung
Diskretisierung	Clustern, zeitliche Diskretisierung, numerische Diskretisierung	• Kontinuierliche, numerische Daten • Klare Verteilungsmuster
Binärisation	One-Hot Encoding, Label Encoding	• Nicht-numerische oder ordinale Daten • Kategorische Variablen mit begrenzten Ausprägungen • Keine hochkardinalen Merkmale
Datenglättung	Gleitender Durchschnitt, Gleitender Median, T4253-Glättung	• Numerische Daten • Zeitreihen mit Erhebungszeitpunkten optimal • Rauschen vorhanden

In Tabelle 3 sind die Aufgaben der Datentransformation dargestellt. Zu jeder Aufgabe werden exemplarisch Vorverarbeitungsverfahren genannt und durch die entsprechenden Anforderungen an die Rohdaten ergänzt. Für Verfahren der Normalisierung, Skalierung und Standardisierung lässt sich die Anforderung numerischer Merkmale mit dem Einhaltungskriterium der Kompatibilität bezüglich Datentyps erklären. Verfahren wie Min-Max-Normalisierung, Z-Wert-Normalisierung oder Dezimalskalierung führen arithmetische Operationen auf Werten durch und benötigen daher numerische Eingabedaten (vgl. Abschnitt 3.2). Die Anforderung, extreme Ausreißer zu vermeiden, ergibt sich aus dem Einflusskriterium hinsichtlich der Datenqualität. Einige Verfahren wie die Min-Max-Normalisierung reagieren sehr empfindlich auf solche Ausreißer, weil diese die Spannweite stark beeinflussen und dadurch alle anderen Werte stark zusammenschieben. Diese Empfindlichkeit beeinträchtigt die Wirksamkeit der Transformation erheblich, auch wenn das Verfahren technisch durchführbar bleibt.

Die Gruppierung der Verteilungstransformationen fordert ebenfalls numerische Daten, was sich aus dem Einhaltungskriterium der Datentypkompatibilität ergibt. Parametrische Verfahren wie Box-Cox-Transformation benötigt spezifisch kontinuierliche Daten (vgl. Abschnitt 3.2). Die Anforderung nach positiven Werten bei Log- und Box-Cox-Transformationen folgt aus der mathematischen Definition des Logarithmus, der nur für positive reelle Zahlen definiert ist. Dies stellt ein absolutes Einhaltungskriterium der Kompatibilität bezüglich Formats dar. Das Einflusskriterium der Datenqualität bildet sich in der Anforderung zur Vermeidung extremer Ausreißer ab, da diese die Transformationsergebnisse verzerren und die Normalisierungswirkung beeinträchtigen können (vgl. Abschnitt 3.2). Ein weitere Einflusskriterium ist das der Dimensionalität, das zur Anforderung nach Merkmalen mit unterschiedlichen Skalen führt, da Verteilungstransformationen insbesondere dann sinnvoll sind, wenn Skalenunterschiede zwischen Merkmalen bestehen. Die Anforderung nach schiefen Verteilungen ergibt sich aus dem Einflusskriterium der Datenverteilung. Verteilungstransformationen zielen darauf ab, schiefe Verteilungen zu linearisieren (vgl. Abschnitt 3.2). Bei bereits normalverteilten Daten könnte die Anwendung sogar kontraproduktiv sein, da sie die ursprüngliche günstige Verteilungsform verschlechtern kann.

Für die Diskretisierung ist die Hauptanforderung, dass die Rohdaten kontinuierlich, numerische Daten enthalten. Dies resultiert aus dem Einhaltungskriterium der Kompatibilität bezüglich Datentyps. Es liegt in der Natur der Aufgabe, die darauf abzielt, den Wertebereich kontinuierlicher Merkmale in diskrete Intervalle oder Kategorien zu überführen. Ohne kontinuierliche Ausgangsdaten ist keine sinnvolle Diskretisierung möglich, wodurch diese Anforderung technisch absolut ist (vgl. Abschnitt 3.2). Aus dem Einflusskriterium der Datenverteilung folgt die Anforderung nach klaren Verteilungsmustern in den Rohdaten. Ein klares Verteilungsmuster in den Rohdaten kann die Anwendbarkeit und potenziell auch die Wirksamkeit bestimmter Diskretisierungsverfahren beeinflussen.

Verfahren der Binärisation sind in erster Linie für kategorische oder ordinale Daten konzipiert, da sie Variablen, die nicht numerisch sind in binäre Repräsentationen überführen (vgl. Abschnitt 3.2). Mit dem Einflusskriterium der Dimensionalität ergibt sich die Anforderung nach begrenzten Ausprägungen und der Vermeidung hochkardinaler Merkmale. Diese Anforderungen sind praktisch motiviert, da hochkardinale kategorische Variablen bei One-Hot-Encoding zu einer exponentiellen Erhöhung der Dimensionalität führen können, wodurch nachgelagerte Algorithmen durch den Fluch der Dimensionalität beeinträchtigt werden (vgl. Abschnitt 3.3).

Für Verfahren der Datenglättung ergibt sich die Anforderung nach numerischen Daten aus dem Einhaltungskriterium der Kompatibilität bezüglich Datentyps. Glättungsalgorithmen nutzen mathematische Operationen wie gleitende Durchschnitte oder Gewichtungsfunktionen (vgl. Abschnitt 3.2) und führt dazu zur absoluten Notwendigkeit dieser Anforderung. Weiter noch resultiert das Einflusskriterium der Zeitreihenstruktur in der Anforderung nach Zeitreihen mit definierten Erhebungszeitpunkten.

Um die Vergleichbarkeit der Datentransformationsverfahren sicherzustellen, wird an dieser Stelle eine weitere Eingrenzung vorgenommen. Neben Aufgaben der Datentransformation, die auf Datenebene ablaufen, werden insbesondere verfahrensspezifische Transformationen betrachtet, da hierzu ein direkter Bedarf im Kontext von Data Mining besteht. Ausgeschlossen werden hingegen Aufgaben, die stark domänenspezifisch sind, gegebenenfalls Expertenwissen erfordern oder deterministische Ergebnisse liefern. Im weiteren Verlauf dieser Arbeit liegt der Fokus daher auf linearen Transformationen sowie auf Verteilungstransformationen. Die entsprechende Operationalisierung der Anforderungen dieser Datentransformationsaufgaben sind in Abbildung 6 dargestellt.

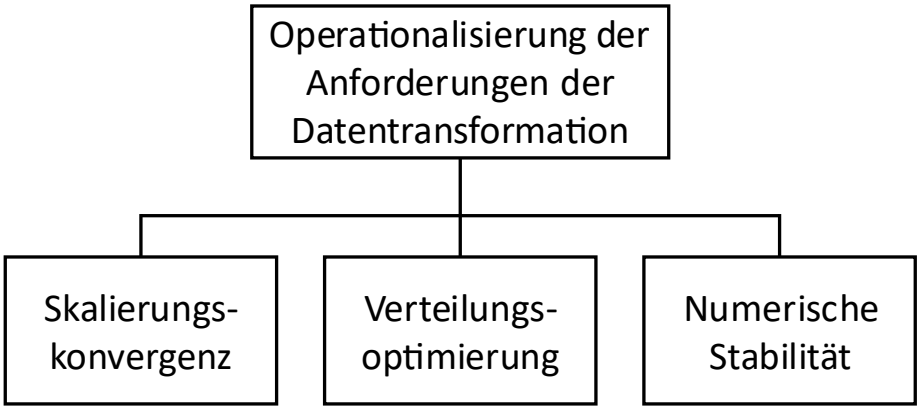

Abbildung 6: Operationalisierung der Anforderungen der Datentransformation

In Abbildung 6 sind die drei Dimensionen Skalierungskonvergenz, Verteilungsoptimierung und numerische Stabilität dargestellt. Diese wurden durch die Analyse der Anforderungen der Datentransformation aus Tabelle 3 abgeleitet. Tabelle 3 zeigt das Problem unterschiedlich ska-

lierter Merkmale in Rohdaten. Dominieren größere Wertebereiche, so führt die Uneinheitlichkeit der Skalierung zwischen Merkmalen insbesondere bei distanzbasierten Verfahren zu Verzerrungen. Datentransformationsverfahren wie Min-Max-Normalisierung und Z-Wert-Normalisierung zielen darauf ab, diese Skalierungsunterschiede zu harmonisieren und eine gleichmäßige Gewichtung aller Merkmale zu ermöglichen (vgl. Abschnitt 3.2). Der Index zur Skalierungsheterogenität (SHI) wird zur Bewertung von Skalierungsunterschieden zwischen Merkmalen in einem Datensatz verwendet. Dieser ist als Adaption des Variationskoeffizienten für Wertebereiche entwickelt, um insbesondere heterogene Skalierungen in Merkmalen quantitativ zu erfassen (vgl. Anhang A). Für jedes Merkmal i wir der Wertebereich als Differenz zwischen Maximum und Minimum bestimmt:

$$Spannweite = \max(X_i) - \min(X_i) \tag{5}$$

Anschließend wird die Standardabweichung dieser Spannweite $\sigma_{Spannweite}$ durch ihren Mittelwert $\mu_{Spannweite}$ geteilt:

$$SHI = \frac{\sigma_{Spannweite}}{\mu_{Spannweite}} \tag{6}$$

Ein SHI nahe Null deutet auf eine homogene Skalierung der Merkmale hin. Als Interpretationsrichtlinie deuten Werte von 0 bis $0{,}3$ auf eine homogene Skalierung hin. Ergänzend bewertet der Streuungsnutzungsgrad (SG) die optimale Ausnutzung des verfügbaren Wertebereichs nach der Transformation. Mathematisch wird der SG als Mittelwert μ der Quotienten aus Standardabweichung des Merkmals σ_i und Spannweite über alle Merkmale i berechnet:

$$SG = \mu\left(\frac{\sigma_i}{\max(X_i) - \min(X_i)}\right) \tag{7}$$

Mit dem Verhältnis $\frac{\sigma_i}{\max(X_i) - \min(X_i)}$ wird gemessen welcher Anteil des Wertebereichs durch die Streuung eines Merkmals genutzt wird. Schließlich wird der Mittelwert über alle Merkmale gebildet, damit die Effizienz über alle Merkmale zu einer Kennzahl aggregiert werden kann.

In Tabelle 3 sind problematische Verteilungseigenschaften wie schiefe Verteilungen und extreme Ausreißer identifiziert worden. Box-Cox-Transformationen und logarithmische Transformationen zielen darauf ab, Verteilungen zu normalisieren. Die Verteilungsoptimierung stellt sich somit als wichtiger Faktor für den Erfolg heraus. Die Normalitätsverbesserung (NV) ist eine Metrik zur Bewertung des Effekts von Datentransformationsverfahren auf die Annäherung der Verteilung der Daten an die Normalverteilung. Die NV quantifiziert die relative Steigerung des p-Werts des Shapiro-Wilk-Test (vgl. Anhang A) nach einer Transformation:

$$NV = \frac{p_{nach} - p_{vor}}{1 - p_{vor}} \tag{8}$$

Hierbei wird der p-Wert vor der Datentransformation als p_{vor} und der p-Wert nach der Datentransformation als p_{nach} bezeichnet. Die Differenz der beiden p-Werte drückt die absolute Veränderung der Normalität aus. Durch das Verhältnis zum maximal möglichen Verbesserungsspielraum $1 - p_{vor}$ ermöglicht diese Metrik den Vergleich verschiedener Verfahren unabhängig vom Ausgangszustand der Daten. Die NV berechnet somit also die relative Verbesserung hinsichtlich der Erreichung einer normalverteilten Struktur. Nimmt die NV hierbei positive Werte an so hat die Datentransformation die Normalität der Daten verbessert, wobei ein NV-Wert nahe eins auf eine nahezu vollständige Normalisierung der Werte hinweist. Ein NV-Wert von null bedeutet, dass keine Veränderung eingetreten ist, während negative Werte auf eine Verschlechterung hindeuten. Zusätzlich neben der NV ergänzt die Symmetrieverbesserung (SV) die Dimension Verteilungsoptimierung. Die SV quantifiziert, in welchem Maße eine Datenverteilung nach der Datenvorverarbeitung symmetrischer geworden ist. Hierbei wird die Ausgangsschiefe berücksichtigt. Die Berechnung erfolgt auf Grundlage der Fisher-Pearson-Skewness-Formel mit der die Asymmetrie einer Verteilung erfasst werden kann (vgl. Anhang A). Die SV ist wie folgt definiert:

$$SV = \frac{|skew_{vor}| - |skew_{nach}|}{|skew_{vor}|} \tag{9}$$

Bei dieser Metrik ist es nicht relevant, ob eine Verteilung links- oder rechtsschief ist, weswegen Beträge verwendet werden, um die Richtungsunabhängigkeit sicherzustellen. Ein positiver Wert zeigt, dass die Schiefe reduziert wurde. Ein Wert von null hingegen bedeutet, dass die Datentransformation keine Wirkung auf die Schiefe hatte. Entsprechend verweist ein negativer Wert auf eine Verschlechterung hin.

Die Anforderung, extreme Ausreißer zu vermeiden, sowie die Notwendigkeit positiver Werte bei bestimmten Transformationen, unterstreichen die Relevanz der numerischen Stabilität. Für die Bewertung dieser wird die Konditionszahl $\kappa(X)$, die ein Maß aus der linearen Algebra ist, genutzt (vgl. Anhang A):

$$\kappa(X) = \frac{\lambda_{\max}(X)}{\lambda_{\min}(X)} \tag{10}$$

Hierbei bezeichnen $\lambda_{\max}(X)$ und $\lambda_{\min}(X)$ die größten beziehungsweise kleinsten Singularwerte der Matrix X. Die Konditionszahl misst inwieweit die Datenstruktur durch Streckung oder Stauchung entlang verschiedener Richtungen beeinflusst wird. Liegt dieser Wert bei annähernd eins deutet dies auf eine gute Konditionierung hin, wohingegen deutlich höhere Werte

auf eine schlechte Konditionierung hindeuten. Ursachen hierfür können gerade starke Unterschiede in den Skalen der Merkmale oder Multikollinearität sein. Es besteht folglich eine nahezu lineare Abhängigkeit zwischen den Spalten von X, was zu numerischer Instabilität führt.

Die Evaluationskriterien aus Abschnitt 4.2 haben gezeigt, dass insbesondere die Anforderungen der nachgelagerten Data-Mining-Algorithmen von großer Bedeutung sind. Eine genaue Betrachtung der verschiedenen Data-Mining-Aufgaben ist somit unerlässlich. Diese Anforderungen können direkt aus der methodischen Funktionsweise und den mathematischen oder strukturellen Annahmen der einzelnen Algorithmen bestimmt werden und sind in Tabelle 4 dargestellt.

Tabelle 4: Anforderungen Data Mining

Aufgabe	**Beispielalgorithmen**	**Anforderungen an Eingabedaten**
Klassifikation	Entscheidungsbäume (ID3, C4.5), k-NN, Neuronale Netze	• Klassenzugehörigkeit (Diskretzielvariable) • Spezif. Datentyp (nominal für C4.5, metrisch für k-NN)
Clusteranalyse	Hierarchische und partitionierende Methoden	• Datentypen, die Distanz-/Ähnlichkeitsmaße erlauben • Numerische Daten oft bevorzugt
Regression	Lineare, polynomiale und schrittweise Regressionsalgorithmen	• Numerische Daten • Linearität zwischen den Variablen • Keine extremen Ausreißer
Wirkzusammenhänge	Apriori	• Nominale oder diskrete Daten • Daten, bei denen gemeinsames Auftreten von Attributen von Interesse ist (Transaktionsbasierte Daten)

Tabelle 4 zeigt die Data-Mining-Aufgaben mit zugehörigen Beispielalgorithmen und den Anforderungen an die Eingabedaten. Die Anforderungen für Klassifikationsalgorithmen leiten sich primär aus dem Einhaltungskriterium der Kompatibilität bezüglich Formats ab. Die grundlegende Voraussetzung für alle überwachten Klassifikationsverfahren ist die Existenz einer diskreten Zielvariable, die die Klassenzugehörigkeit angibt (vgl. Abschnitt 2.4). Ohne Zielvariable kann keine überwachte Lernprozedur durchgeführt werden, weswegen diese Anforderung

zwingend erforderlich ist. Das Einhaltungskriterium der strukturellen Verfügbarkeit wird hier durch die Notwendigkeit einer klar definierten Label-Struktur konkretisiert. Das Einhaltungskriterium der Kompatibilität bezüglich Datentyps zeigt sich in den spezifischen Anforderungen an die Eingabedaten. Hierzu sind beispielsweise baumbasierte Verfahren wie ID3 oder C4.5 sowie probabilistische Ansätze wie Naive Bayes für die Verarbeitung kategorialer Informationen konzipiert und setzen daher nominale oder diskretisierte Daten voraus (vgl. Abschnitt 2.4). Im Gegensatz dazu erfordern distanzbasierte Verfahren wie k-NN kontinuierliche, metrische Attribute, da sie auf Distanzberechnungen im Merkmalsraum beruhen. Diese Anforderung folgt aus dem Einhaltungskriterium der Kompatibilität bezüglich Datentyps. Distanzmetriken wie die euklidische Distanz sind nur auf numerischen Daten definiert. Das Einflusskriterium der Datenverteilung zeigt sich in der Sensitivität distanzbasierter und verteilungsannahmebasierter Verfahren gegenüber Skalierung und Unterschieden hinsichtlich Bei k-NN können beispielsweise Merkmale mit größeren Wertebereichen die Distanzberechnung dominieren, während bei neuronalen Netzen unterschiedliche Skalen zu instabilen Gewichtungen führen können (vgl. Abschnitt 2.4).

Für die Aufgabe der Clusteranalyse unterscheiden sich die Anforderungen grundlegend von überwachten Algorithmen durch das Fehlen einer Zielvariable. Das Einhaltungskriterium der Formatkompatibilität zeigt sich hier konkret in der zwingenden Abwesenheit von Klassenlabels, weil die Gruppierung allein auf intrinsischen Datenstrukturen beruhen muss. Das Einhaltungskriterium der Kompatibilität bezüglich des Datentyps ergibt sich aus der Notwendigkeit, Ähnlichkeits- oder Distanzmaße zwischen Objekten zu berechnen (vgl. Abschnitt 2.4).

Für Algorithmen der Regression leitet sich die Anforderung kontinuierlicher, numerischer Zielvariable aus dem Einhaltungskriterium der Kompatibilität bezüglich des Formats ab. Diese Anforderung ist notwendig, da Regression darauf abzielt quantitative Werte vorherzusagen oder zu schätzen (vgl. Abschnitt 2.4). Das Einhaltungskriterium der Kompatibilität bezüglich Datentyps variiert je nach spezifischem Regressionsverfahren. Das Einflusskriterium der Datenverteilung und der Datenqualität resultiert in den Anforderungen zur Linearität und keiner extremen Ausreißer. Lineare Regressionsverfahren sind sensitiv gegenüber der Skalierung der Eingabemerkmale, da unterschiedliche Skalen zu numerischen Instabilitäten bei der Parameteroptimierung führen können. Bestimmte Verteilungscharakteristika der Zielvariable können die Modelleistung beeinträchtigen, insbesondere bei Vorhandensein von Heteroskedastizität oder nicht-linearen Zusammenhängen (vgl. Abschnitt 2.4).

Die algorithmusspezifischen Anforderungen verdeutlichen die Diskrepanz zwischen den Eigenschaften der Rohdaten und den Voraussetzungen datenanalytischer Verfahren. Diese Diskrepanz muss durch gezielte Datenvorverarbeitung überwunden werden, sodass eine algo-

rithmische Kompatibilität hergestellt werden kann. Nur so kann eine erfolgreiche Wissensentdeckung ermöglicht werden. Die Erfüllung dieser Anforderungen und die Operationalisierung dieser ist in Abbildung 7 dargestellt.

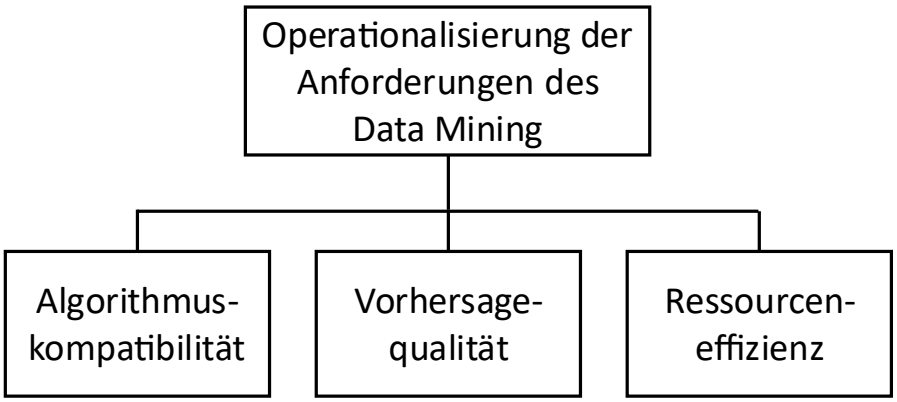

Abbildung 7: Operationalisierung der Anforderungen des Data Mining

In Abbildung 7 sind die Dimensionen Algorithmuskompatibilität, Vorhersagequalität und Ressourceneffizienz dargestellt. Die in Tabelle 4 gelisteten Anforderungen verdeutlichen, wie spezifisch sie auf die jeweilige Data-Mining-Aufgabe zugeschnitten sind. Beispielsweise sind diskrete Zielvariablen für Klassifikation, numerische Daten für Regression und die Erlaubnis von Distanzmaßen für Clustering gefordert. Dies verdeutlicht die Notwendigkeit einer genauen Anpassung der Eingabedatencharakteristika an die algorithmischen Voraussetzungen. Die Dimension der Algorithmuskompatibilität wird stark von der Wahl des Data-Mining-Algorithmus beeinflusst. Eine fundierte Bewertung ist nur möglich, wenn mehrere Einflussfaktoren berücksichtigt werden. Hierzu können die drei Dimensionen und die entsprechenden Metriken (4) und (6) bis (10) herangezogen werden. Diese Metriken ermöglichen eine differenzierte Bewertung der Eignung vorverarbeiteter Daten in Bezug auf Skalierung, Verteilung und numerische Stabilität. Dies sind Aspekte, die die Leistungsfähigkeit von Data-Mining-Algorithmen wesentlich mitbestimmen.

Die Dimension der Vorhersagequalität bewertet übergeordnet den Erfolg der Datenvorverarbeitungskette und des Data-Mining-Schrittes durch direkte Messung der algorithmischen Performanz (vgl. Abschnitt 2.4). Alle Data-Mining-Anforderungen in Tabelle 4 zielen maßgeblich auf die Verbesserung der Modellleistung ab. Spezifische Anforderungen wie lineare Beziehungen für Regression, keine extremen Ausreißer oder bestimmte Datentypen stellen Mittel zur Erreichung der optimalen Vorhersagequalität dar.

Um eine stabile und effiziente Ausführung des Data-Mining-Algorithmus zu gewährleisten, wird die Dimension Ressourceneffizienz berücksichtigt. Dateneigenschaften, die den Anforderungen des Data Mining nicht genügen, können numerische Instabilitäten, Konvergenzprobleme oder einen übermäßigen Ressourcenverbrauch verursachen. Hierin wird die Recheneffizienz durch Messung der Trainingszeit und des Speicherverbrauchs quantifiziert und bewertet die

praktische Anwendbarkeit der Vorverarbeitungsergebnisse. Die Ressourceneffizienz stellt sicher, dass nicht nur die Datenqualität, sondern auch die praktische Durchführbarkeit der nachgelagerten Analysen berücksichtigt wird.

Metriken

Aus den bisherigen Erkenntnissen in diesem Kapitel, lassen sich die Metriken für den Vergleich in eine strukturierte Übersicht überführen. Die Metriken, die im Rahmen des isolierten Vergleichs den Erfolg eines Datenvorverarbeitungsschrittes im Kontext von Data Mining quantifizieren, sind in Abbildung 8 dargestellt.

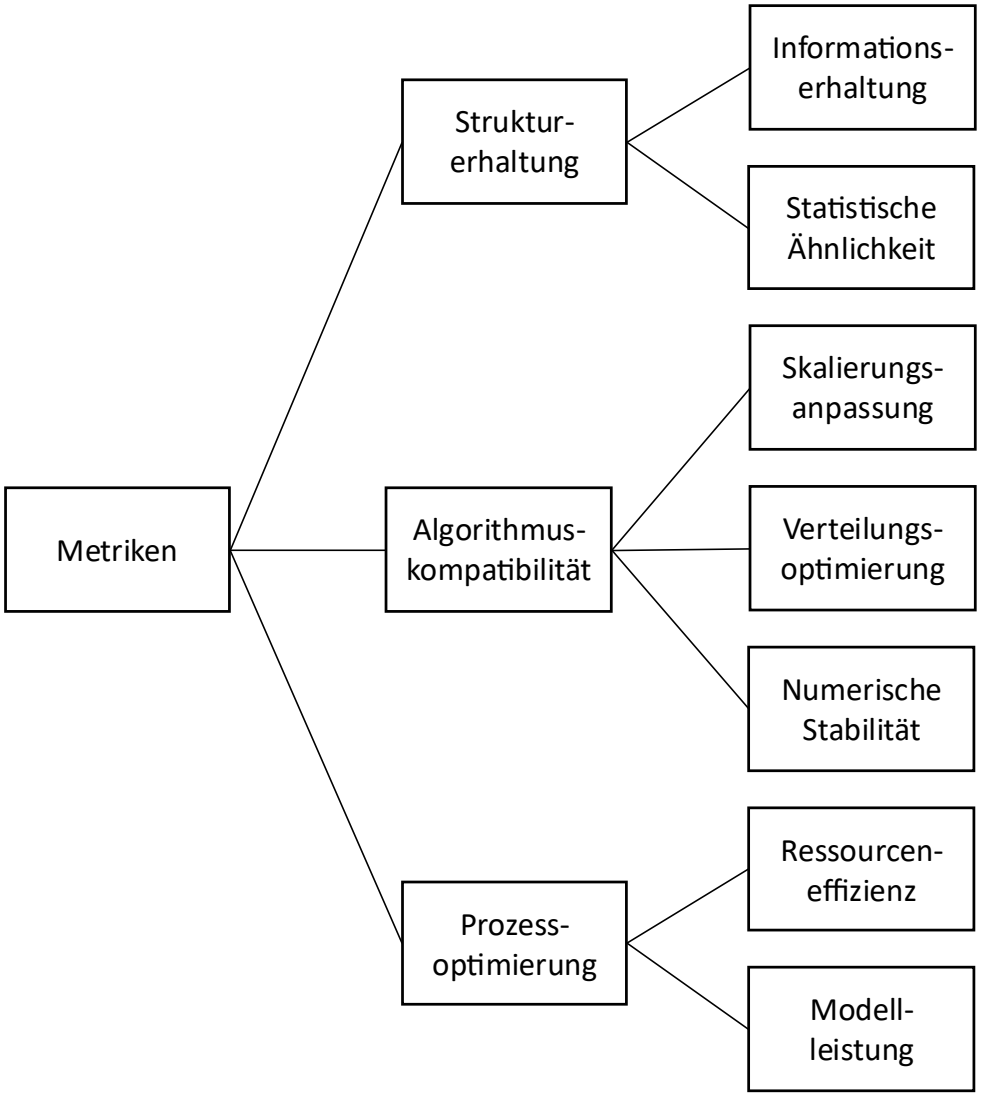

Abbildung 8: Metriken für den isolierten Vergleich der Datentransformation und Datenreduktion

Abbildung 8 zeigt ein Baumdiagramm zur Einordnung der Metriken. Ausgehend vom Wurzelknoten Metriken verzweigt sich die Struktur in die drei Hauptkategorien Strukturerhaltung, Algorithmuskompatibilität und Prozessoptimierung. Strukturerhaltung umfasst Metriken zur Informationserhaltung sowie statistischen Ähnlichkeit. Diese Metriken prüfen, ob wesentliche Datencharakteristika trotz Datentransformation und Datenreduktion bewahrt bleiben. Hierunter

kann die Informationserhaltung direkt über (2) gemessen werden (vgl. Anhang B). Der Knoten zur statistischen Ähnlichkeit kann durch drei komplementäre Metriken operationalisiert werden. Hierzu gehört die Korrelationserhaltung, die die Stabilität linearer Merkmalsbeziehungen bewertet. Auch die Cluster-Trennbarkeit sowie die Varianzerhaltung in den ersten k Hauptkomponenten, die prüft, ob dominante Varianzquellen trotz Datenreduktion erhalten bleiben, lassen sich hier einordnen (vgl. Anhang A). Diese Zusammenstellung deckt sowohl globale als auch lokale sowie dimensionsspezifische Aspekte zur Struktur ab. Die Kategorie Algorithmuskompatibilität gliedert sich in Skalierungsanpassung, Verteilungsoptimierung und numerische Stabilität. Die Skalierungsanpassung integriert die Skalierungshomogenität (6) und den Streuungsnutzungsgrad (7), da viele Data-Mining-Algorithmen homogen skalierte Merkmale benötigen. Die Verteilungsoptimierung bündelt mit Normalitäts- (8) und Symmetrieverbesserung (9) Metriken, die für Verfahren wie lineare Regression relevant sind, welche Normalverteilungsannahmen treffen. Die numerische Stabilität umfasst die Konditionszahl (10) und die Merkmalsinterdependenz (4), die Probleme der Multikollinearität in linearen Modellen aufdeckt. Diese Gruppierung spiegeln die unterschiedlichen Datenvoraussetzungen der Algorithmen wider. Weiter umfasst die Kategorie Prozessoptimierung die Metriken, die sich unter Ressourceneffizienz und Modellleistung einordnen lassen. Innerhalb der Ressourceneffizienz werden hierzu die Rechenzeit und der Speicherbedarf hinsichtlich des Data-Mining-Schrittes zur Quantifizierung der praktischen Umsetzbarkeit genutzt. Die Modellleistung misst schließlich als finale Zielgröße die Modellgüte.

Die Zuordnung der Metriken folgt somit einem kaskadierenden Prinzip. Die Metriken reichen von der Bewahrung einzelner Merkmale über die Kompatibilität zu Algorithmen bis hin zur übergeordneten Bewertung des gesamten Prozesses von Datenvorverarbeitung und Data Mining. Weiter werden für den kombinierten Vergleich die Wechselwirkungen und Reihenfolgeeffekte zwischen der Datentransformation und Datenreduktion untersucht. Die entsprechende Übersicht zu Einordnung der Metriken ist in Abbildung 9 dargestellt.

In Abbildung 9 ist ein Baumdiagramm zur Einordnung der Metriken für den kombinierten Vergleich der Datentransformation und Datenreduktion dargestellt. Die Struktur entspricht grundsätzlich der Darstellung in Abbildung 8, wobei der Knoten für Verteilungsoptimierung entfallen ist. Im Rahmen des kombinierten Vergleichs werden ausschließlich solche Metriken berücksichtigt, deren Aussagekraft in beiden Vorverarbeitungsschritten erhalten bleibt. Datentransformationen verbessern die statistischen Eigenschaften einzelner Merkmale. Für die Kategorie der Verteilungsoptimierung sind hierzu Normalität und Symmetrie berücksichtigt worden. Solche Optimierung sind jedoch nur im ursprünglichen Merkmalsraum sinnvoll interpretierbar. Nach der Datenreduktion liegen Daten in einem neuen orthogonalen Raum vor, dessen Achsen lineare Kombinationen der Originalmerkmale darstellen (vgl. Abschnitt 3.4).

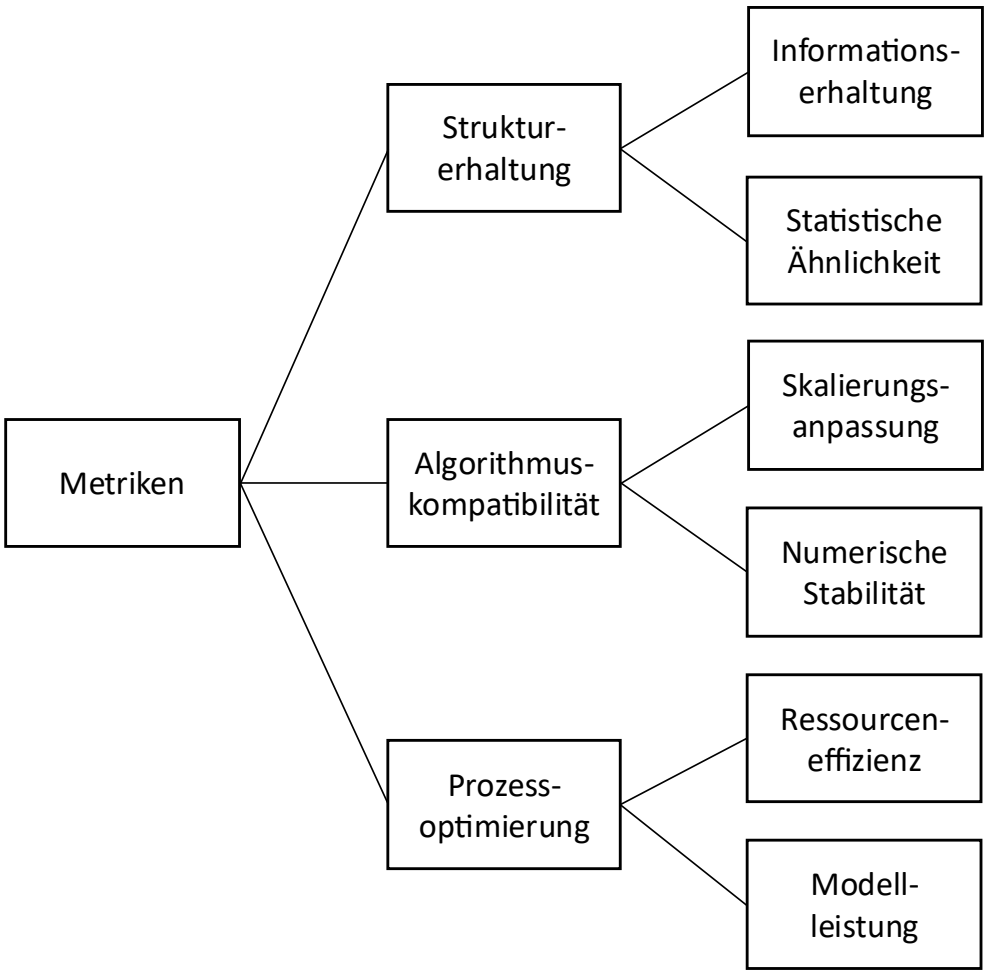

Abbildung 9: Metriken für den kombinierten Vergleich der Datentransformation und Datenreduktion

Folglich sind Metriken wie die Normalitätsverbesserung oder Symmetrieverbesserung nach einer Datenreduktion nicht mehr konsistent interpretierbar und werden im kombinierten Vergleich ausgeschlossen.

Zur besseren Übersicht werden die in diesem Abschnitt den entsprechenden Kategorien zugeordneten Metriken, die in den Abbildungen 8 und 9 als Endknoten dargestellt sind, in einer Tabelle zusammengefasst, die im Anhang B zu sehen ist.

4.4 Entwicklung eines Vergleichrahmens zum Vergleich von Datentransformation- und Datenreduktionsverfahren

In diesem Kapitel wird ein experimenteller Vergleichsrahmen entwickelt, der eine systematische Analyse und Bewertung von Datentransformations- und Datenreduktionsverfahren ermöglicht. Ziel ist es, deren isolierten sowie kombinierten Effekt auf Datenqualität, Algorithmusperformanz und Ressourceneffizienz zu erfassen. Eine Modifikation des Vergleichsrahmen ermöglicht es weiter noch Reihenfolgeeffekte und Synergien zwischen den Verfahrens-

schritten zu untersuchen. Die in Abschnitt 4.3 entwickelten Metriken werden in den allgemeinen Vergleichsrahmen implementiert und ermöglichen einen methodischen Experimentierablauf.

Für den Vergleich der Datentransformations- und Datenreduktionsverfahren selbst müssen je nach Anwendungsfall und Analyseziel die Rohdaten selbst zunächst vorbereitet werden. Nach dem Vorgehensmodell von Fayyad et al. (1996) wird aus den Rohdaten zunächst ein Zieldatensatz ausgewählt, der anschließend von Rauschen und Inkonsistenzen bereinigt werden muss (vgl. Abschnitt 2.3). Der in diesem Kapitel entwickelte Vergleichsrahmen setzt jedoch erst bei der vierten Phase zur Datenreduktion und Projektion an. Aus diesem Grund werden die Ausgangsdaten im Folgenden, trotz bereits erfolgter Auswahl und Vorverarbeitung, weiterhin als Rohdaten bezeichnet. Der grundsätzliche Ablauf für den eigentlichen Vergleich ist in Abbildung 10 zu sehen.

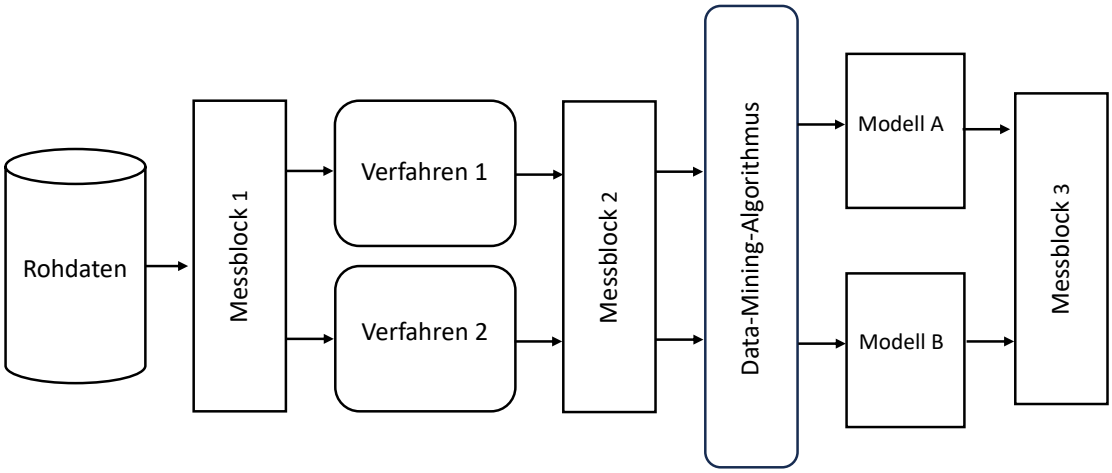

Abbildung 10: Grundsätzlicher Ablauf des isolierten Vergleichs

Der in Abbildung 10 dargestellte Ablauf beginnt mit einem Rohdatensatz, der zunächst einen ersten Messblock durchläuft. Dieser ist notwendig, da einige Metriken zur Berechnung einen direkten Bezug zu den ursprünglichen Eigenschaften der Rohdaten benötigen. Anschließend werden die Rohdaten in zwei getrennten Strängen jeweils mit unterschiedlichen Vorverarbeitungsverfahren bearbeitet. Die hierbei entstehenden vorverarbeiteten Daten stehen nicht im Fokus des Vergleichs und werden aus diesem Grund nicht persistent gespeichert. Stattdessen folgt unmittelbar der zweite Messblock, der sich auf die vorverarbeiteten Daten bezieht. Einerseits dient dieser zweite Messblock der Vervollständigung jener Metriken, die bereits im ersten Messblock beginnen. Andererseits dient der zweite Messblock zur Erfassung zusätzlicher Metriken, die nur auf Basis der vorverarbeiteten Daten vergleichende Aussagen über beide angewandten Verfahren treffen können. Im nächsten Schritt wird auf beide unterschiedlich vorverarbeiteten Datensätze derselbe Data-Mining-Algorithmus angewendet. Dies führt zu zwei un-

terschiedlichen Modellen, die anschließend in den dritten Messblock überführt werden. Messblock 3 erfasst gezielt Metriken, die sich auf die Güte der resultierenden Modelle beziehen (vgl. Abschnitt 2.4).

Abschnitt 4.3 zeigt, dass die Metriken der Kategorien Informationserhaltung, statistische Ähnlichkeit und Verteilungsoptimierung sowohl auf den Rohdaten als auch auf den vorverarbeiteten Daten basieren und daher in Messblock 1 und Messblock 2 Anwendung finden. Darüber hinaus wird deutlich, dass die Metrik Informationserhaltung sowie die Metriken der Kategorie Skalierungsanpassung ausschließlich mit den vorverarbeiteten Daten arbeiten und somit Messblock 2 zugeordnet werden können. Die Metriken zur Prozessoptimierung lassen sich schließlich Messblock 3 zuordnen. Die Zuordnungen der Metriken zu den jeweiligen Messblöcken wird in Vergleichsblöcken zusammengeführt, wie in Abbildung 11 dargestellt.

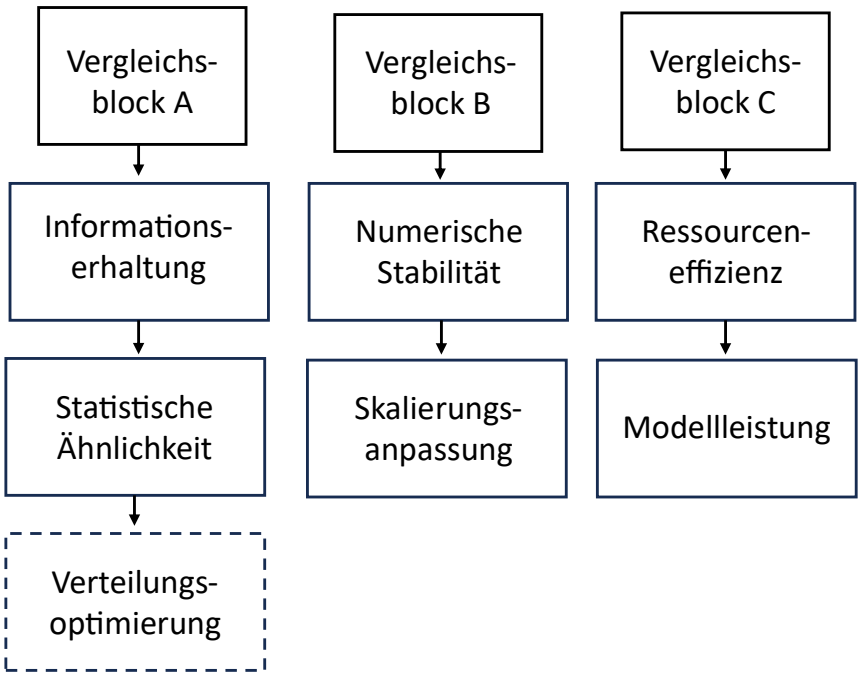

Abbildung 11: Vergleichsblöcke und zugehörige Kategorien

Abbildung 11 zeigt die drei Gruppen der Vergleichsblöcke. Vergleichsblock A umfasst die Metriken, die in Messblock 1 und Messblock 2 zum Einsatz kommen. Die zugehörigen Kategorien sind dort aufgeführt, wobei die Kategorie Verteilungsoptimierung in einem gestrichelten Rahmen dargestellt ist. Dies verweist darauf, dass die hierin enthaltenen Metriken im Rahmen des kombinierten Vergleichs entfallen, da sie an dieser Stelle nicht relevant sind (vgl. Abschnitt 4.3). Vergleichsblock B enthält Metriken, die ausschließlich im zweiten Messblock verwendet werden. Vergleichsblock C bezieht sich auf die Metriken, die im Data-Mining-Schritt zur Bewertung der Modelle eingesetzt werden.

Im Rahmen der Vergleichsbedingungen ist das Analyseziel ein wichtiger Einflussfaktor, der direkt mit der späteren Data-Mining-Aufgabe verbunden ist. Die Datenvorverarbeitung bereitet die Daten so vor, dass sie vom gewählten Algorithmus sinnvoll genutzt werden können. Jeder

Algorithmus bringt spezifische Anforderungen an Struktur, Format und Qualität der Eingangsdaten mit (vgl. Abschnitt 4.3). Die Wirksamkeit der Vorverarbeitung zeigt sich daran, wie stark sie die Leistung des eingesetzten Verfahrens unterstützt oder verbessert. Je nach Zielsetzung kann eine sinnige Interpretation der Metriken erfolgen, nach denen die Verfahren im Vergleich eingeordnet werden. Die Passung zwischen Daten und Analysezweck ergibt sich aus der erfolgreichen Ausrichtung der Daten auf dieses Ziel. Fachliche Anforderungen aus dem jeweiligen Anwendungsgebiet spielen ebenfalls eine Rolle, um sicherzustellen, dass die Resultate nicht nur sinnvoll interpretierbar, sondern auch praktisch nutzbar sind. Mit den Einflusskriterien (s. Tabelle 1) zeigt sich der Anwendungskontext folglich als ein weiterer relevanter Aspekt. Dieser betrifft sowohl technische als auch inhaltliche Rahmenbedingungen. Die Art der Vorverarbeitung, ob in Echtzeit oder in größeren Datenpaketen, beeinflusst die Anforderungen, Schnelligkeit und Skalierbarkeit. In diesem Zusammenhang ist die Qualität der Daten von besonderem Interesse (vgl. Abschnitt 2.2). Die Eignung der Daten für den Zweck der Analyse ist ein direktes Ergebnis der erfolgreichen Anpassung der Daten an die domänenspezifischen Anforderungen. Ein weiterer Einflussfaktor betrifft die technischen Ressourcen. Begrenzungen durch verfügbare Rechenleistung, Laufzeit und Speicher bestimmen, welche Verfahren praktisch umsetzbar sind. Überschreiten Verfahren der Datentransformation und Datenreduktion die vorhandenen Ressourcen, so sind sie ungeeignet für den Vergleich. Neben der Hardware beeinflusst auch die technische Infrastruktur die Auswahl geeigneter Verfahren. Der Vergleichsrahmen ist unter der Voraussetzung entwickelt worden, dass die Rohdaten in relationalen Datenbanken vorliegen. Die Leistungsfähigkeit SQL-basierter Operationen kann hierbei die praktische Umsetzbarkeit beeinflussen, während die tabellarische Struktur solcher Datenbanken bestimmte Vorverarbeitungsschritte begünstigt (vgl. Abschnitt 2.2). Die Einflussfaktoren, der Ablauf des Vergleichs und die Vergleichsblöcke sind als gesamter, strukturierter Vergleichsrahmen in Abbildung 12 dargestellt.

Abbildung 12 zeigt die auf den Vergleichsrahmen wirkenden Einflussfaktoren Analyseziel, Anwendungskontext und verfügbare Rechenressourcen.

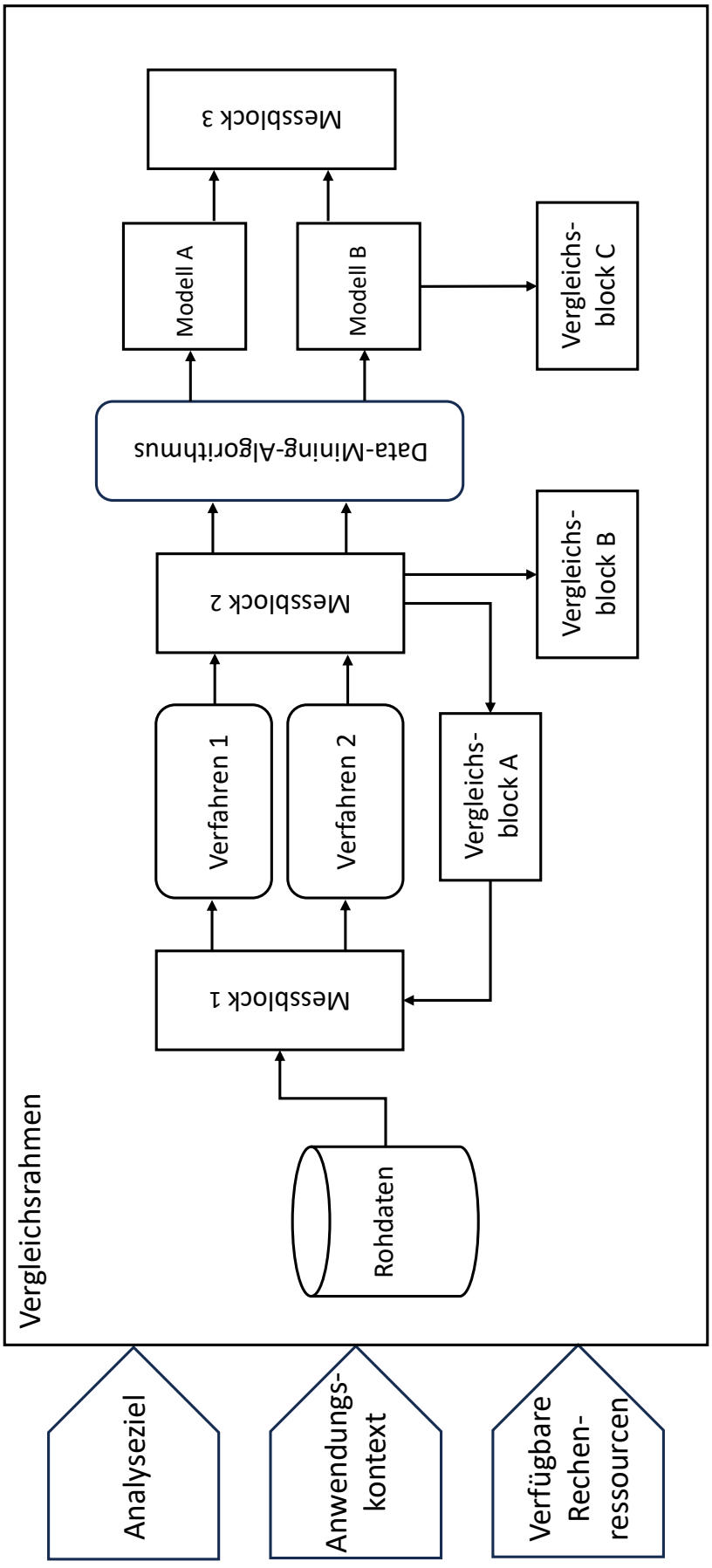

Abbildung 12: Vergleichsrahmen für isolierte Vergleiche von Datentransformations- und Datenreduktionsverfahren

Der Ablauf des Vergleichs selbst ist innerhalb des Vergleichrahmens dargestellt. Hierbei ist zu beachten, dass im isolierten Vergleich entweder ausschließlich Verfahren der Datenreduktion oder ausschließlich Verfahren der Datentransformation angewendet werden. Parallel zum Ablauf sind die Vergleichsblöcke abgebildet, die mit den jeweils relevanten Messblöcken verknüpft sind. Aus diesen werden die entsprechenden Messwerte übernommen. Die den Vergleichsblöcken zugeordneten Metriken sind in Abbildung 11 und im Anhang B aufgeführt. Für den kombinierten Vergleich erfordert der Ablauf lediglich eine angepasste Vorverarbeitung zwischen Messblock 1 und Messblock 2. Um die vollständige Wiederholung des gesamten Vergleichsrahmens zu vermeiden, zeigt Abbildung 13 den entsprechend angepassten Ablauf des Vergleichs.

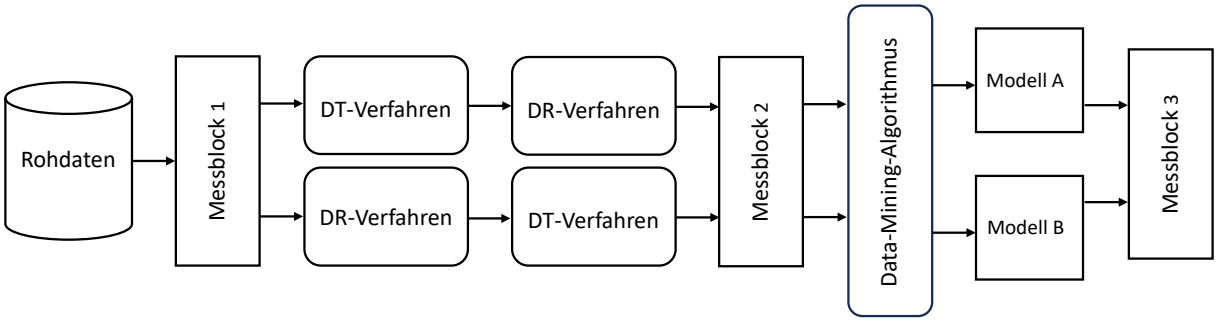

Abbildung 13: Grundsätzlicher Ablauf des kombinierten Vergleichs

Abbildung 13 zeigt die angepasste Vorverarbeitung zwischen Messblock 1 und Messblock 2. In der dargestellten Verzweigung wird entweder zunächst ein Datentransformationsverfahren und anschließend ein Datenreduktionsverfahren angewendet oder genau umgekehrt. Alle übrigen Elemente entsprechen dem Ablauf des isolierten Vergleichs.

Herstellung der Vergleichbarkeit

Die Herstellung der Vergleichbarkeit zwischen unterschiedlichen Aufgaben der Datentransformation muss an dieser Stelle genauer bedacht werden. Gerade die Vielseitigkeit der Datentransformation, deren Aufgaben nicht zwangsläufig die gleiche Art der Transformation vornehmen oder auf identische Ziele ausgerichtet sind, erfordert diese Differenzierung. Wie in Kapitel 3.2 erläutert, lassen sich die Aufgaben der Datentransformation grundsätzlich in die zwei Hauptkategorien zu verfahrensunabhängiger Transformation und verfahrensspezifische Transformation unterteilen. Es ist hier nicht sinnvoll Datentransformationsverfahren zu vergleichen, die fundamental unterschiedliche Aufgaben erfüllen. Weiter noch laufen einige Aufgaben der Datentransformation auf Datenebene ab und nehmen somit strukturelle und syntaktische Anpassungen an den Datensatz vor. Andere Aufgaben wiederum laufen auf Informationsebene und lassen hierdurch teilweise keine zielführenden Vergleichsstudien zu. Letztlich ist es häufig nicht möglich die Datenqualität und die Nützlichkeit oder Relevanz eines verarbeiteten

Datensatzes zu messen. Weiter können diejenigen Aufgaben ausgeschlossen werden, die deterministische Verfahren nutzen. Deren Ergebnisse sind mathematisch eindeutig wie die Umrechnung von Einheiten oder die Berechnung von Aggregationswerten wie Summen oder Mittelwerten. Diese liefern keine vergleichbare methodische Bandbreite und ihre Ausführung somit keinen Spielraum für alternative Ansätze. Zuletzt sind Transformationen, die besonders domänenspezifisch sind, wie die Generalisierung von Städten zu Bundesländern, für einen generischen Vergleich ungeeignet. Ihre Bewertung hängt stark von dem fachlichen Kontext ab und lässt sich nicht durch objektive, universelle Metriken erfassen. In Abschnitt 4.3 wurde die Einschränkung auf lineare Transformationen und Verteilungstransformationen festgelegt. Im Folgenden wird diese Entscheidung weiter begründet. Hierzu müssen die folgenden Punkte von den zu betrachtenden Datentransformationsaufgaben adressiert werden:

- Die Aufgaben bieten zwischen ihren Verfahren unterschiedliche Ergebnisse an
- Die Ergebnisse der Verfahren sind messbar
- Die Aufgaben sind generisch genug, um generalisierbare Aussagen zu ermöglichen

Durch diese Fokussierung kann sichergestellt werden, dass der Vergleich auf Verfahren beschränkt bleibt, die sowohl methodische Diversität als auch eine klare Zuordnung zu evaluierbaren Metriken bieten.

Während bei der Datentransformation durch die Aufgabenvielfalt und unterschiedlichen Zielsetzungen eine sorgfältige Auswahl und Abgrenzung notwendig ist, stellt sich die Situation bei der Datenreduktion anders dar. Die Aufgaben der Datenreduktion verfolgen ein weitgehend gemeinsames Ziel. Dieses besteht in erster Linie darin, die Datenmenge zu verringern ohne wesentliche Informationen zu verlieren.

5 Exemplarische Anwendung des Vergleichsrahmens im Wissensentdeckungsprozess

Das vorliegende Kapitel beschäftigt sich mit der praktischen Anwendung des entwickelten Vergleichsrahmens und der Bewertung der Vorverarbeitungsverfahren im Rahmen einer Fallstudie. Aufbauend auf den in Kapitel 4.3 entwickelten Metriken wird untersucht, wie sich unterschiedliche Datentransformations- und Datenreduktionsverfahren in der Praxis bewähren und welche Auswirkungen sie auf nachgelagerte Data-Mining-Prozesse haben. Insbesondere ist von Interesse, ob die Reihenfolge der Vorverarbeitungsschritte einen messbaren Einfluss auf die Qualität der Endergebnisse ausübt. Hierzu wird der systematische Vergleich zwischen isolierten Einzelverfahren und verschiedenen Kombinationen durchgeführt. Wie für diese Arbeit in Kapitel 2.3 festgelegt, orientiert sich die exemplarische Anwendung methodisch am Vorgehensmodell nach Fayyad et al. (1996). Die entsprechenden Phasen werden im Verlauf dieses Kapitels mehrfach referenziert.

5.1 Einführung des Fallbeispiels

Als Fallbeispiel für die vorgenommene exemplarische Anwendung wird der NASA Turbofan Engine Degradation Datensatz (Saxena und Goebel 2008) verwendet. Da reale Systeme jedoch häufig nicht ausreichend instrumentiert sind, um die für prädiktive Analysen benötigte Daten in ausreichender Qualität und Dichte zu erfassen, bieten synthetisierte Datensätze eine geeignete Grundlage. Die generierten Daten wurden gezielt für einen Datenwettbewerb erstellt. Zur Nachbildung realer Herausforderungen ist ein öffentlich zugängliches Simulationsmodell bereitgestellt. Dadurch kann die Forschung, ohne auf schwer zugängliche, reale Daten angewiesen zu sein, den bereitgestellten Datensatz für die Weiterentwicklung oder den Vergleich von Methoden nutzen. Dieser Datensatz beinhaltet die Aufzeichnung der resultierenden Sensorwerte eines in MATLAB und Simulink implementierten Tools, das ein realistisches großes kommerzielles Stahltriebwerk simulieren kann. Die Daten bilden den gesamten Fehlerverlauf bis zum Ausfall ab und können für die prädiktive Instandhaltung von Nutzen sein. Es werden insgesamt 21 Sensoren erfasst und der Ausgangspunkt für die Degradation eines jeden Moduls wird zufällig gewählt, sodass Fertigungs- und Montagevariationen modelliert sind. Die generierten Daten, liegen als zeitliche Sequenzen von Sensormessungen pro Triebwerk vor, hierunter beispielsweise Temperatur, Drücke und Drehzahlen. Die Restnutzungsdauer (RUL) ist als klare Zielvariable definiert. Zusätzlich sind Rauschquellen integriert, um die Daten realistisch zu gestalten und erfordern somit eine sorgfältige Vorbereitung. Mit insgesamt 26 Merkmalen ist der Datensatz hinreichend komplex, um von Verfahren der Datenreduktion zu profi-

tieren. Darüber hinaus messen die Sensoren in unterschiedlichen Einheiten, weshalb die Anwendung von Verfahren der Datentransformation unverzichtbar ist, um Verzerrungen zu vermeiden.

Für die exemplarische Anwendung wird der Datensatz FD001 verwendet. Dieser beinhaltet ein Trainingsdatensatz, einen Testdatensatz sowie den realen RUL-Wert für jedes Triebwerk. Jede Zeile des Datensatzes repräsentiert einen Zustand eines Triebwerks zu einem bestimmten Zeitpunkt. Dieser Datensatz nutzt einen Fehlermodus und eine Betriebsbedingung, wodurch die spezifischen Auswirkungen der zu untersuchenden Vorverarbeitungsverfahren direkt geprüft werden können, ohne dass Ergebnisse durch unterschiedliche Betriebsbedingungen oder Fehlermodi verfälscht werden. FD001 umfasst 100 Verläufe von Triebwerken, die somit eine statistisch aussagekräftige Grundlage für die Vorverarbeitung bieten. Die experimentelle Umgebung wird konstant gehalten und nur die Vorverarbeitungsverfahren ändern sich, wodurch eine kontrollierte Variation ermöglicht wird. Insbesondere wäre bei einer Variation des Data-Mining-Algorithmus nicht mehr eindeutig nachvollziehbar, ob die Metriken des Vergleichsblocks C des entwickelten Vergleichsrahmens dem Vergleich der Data-Mining-Algorithmen oder der Datenvorverarbeitung dienen. Dies erlaubt eine interne Validität der Untersuchung und macht die Ergebnisse der verschiedenen Verfahrenskombinationen direkt vergleichbar. Der zugrundeliegende Datensatz sowie der verwendete Data-Mining-Algorithmus bleiben konstant, um eine einheitliche Vergleichsbasis zu schaffen und weitere externe Einflüsse zu minimieren.

Verfahrensauswahl

Der Datensatz FD001 zeichnet sich durch eine Vielzahl unterschiedlicher Sensormessungen aus und besitzt hiermit eine hohe Dimensionalität und heterogene Skalenstruktur. Für die Datentransformation wurde die Z-Wert-Normalisierung und die Min-Max-Normalisierung ausgewählt. Diese Verfahren adressieren die uneinheitlichen Sensorskalen, die im Datensatz FD001 durch unterschiedliche physikalische Größen und Betriebseinstellungen entstehen. Data-Mining-Algorithmen, die insbesondere auf Distanzmaßen oder Matrixoperationen beruhen, setzen eine vergleichbare Skalenordnung voraus (vgl. Abschnitt 4.3). Die Auswahl beider Methoden erlaubt den Vergleich zwischen einer standardisierten Transformation auf Basis der Streuung und einer skalenbasierten Normierung, wodurch gerade die Unterschiede hinsichtlich Datenstruktur analysiert werden können. Im Bereich der Datenreduktion werden die PCA und der FastICA eingesetzt. Während die PCA die Daten in einen neuen Vektorraum projiziert, was bei hochdimensionalen Zeitreihendaten von Vorteil ist, zielt FastICA hingegen auf die Entmischung der Daten durch Maximierung der statistischen Unabhängigkeiten der Komponenten.

5.2 Durchführung der exemplarischen Anwendung

Die exemplarische Untersuchung mit diesem Datensatz betrachtet die Wechselwirkungen zwischen Datenvorverarbeitungsverfahren der Datenreduktion- und Datentransformation innerhalb des kompletten Wissensentdeckungsprozesses. Der methodische Ansatz basiert auf der Anwendung des entwickelten Vergleichsrahmens innerhalb eines nach dem Vorgehensmodel nach Fayyad et al. (1996) organisierten Wissensentdeckungsprozesses. Hierzu wird die Phase der Transformation wie bereits erläutert unter kontrollierten Rahmenbedingungen variiert.

5.2.1 Vorbereitung des Wissensentdeckungsprozesses

Der folgende Abschnitt befasst sich zunächst mit Phase 1 des Vorgehensmodells nach Fayyad et al. (1996). Es wird das Datenverständnis und der fachliche Kontext geklärt, sowie ein Analyseziel definiert. Die Anwendungsdomäne ist die industrielle Zustandsüberwachung von Maschinen. Das Ziel ist die frühzeitige Erkennung des Verschleißprozesses. Im konkreten Projektkontext steht die prädiktive Instandhaltung im Vordergrund. Es soll die Prognose des verbleibenden RUL eines Systems ermittelt werden mit dem übergeordneten Ziel Wartungen effizient und dem vorliegenden Zustand entsprechend durchzuführen. Da die RUL-Werte nicht direkt im FD001- Datensatz vorliegen, müssen sie aus den Lebensdauersequenzen für jedes Triebwerk berechnet werden. Für den Datensatz ist eine hohe Dimensionalität und unterschiedliche Anfangszustände der Triebwerke, sowie betriebsbedingte Variationen und sensorisches Rauschen charakteristisch. Dementsprechend müssen die Daten vor dem eigentlichen Vergleich zunächst vorbereitet werden. Diese Vorbereitung wurde in RapidMiner Studio durchgeführt und ist in Abbildung 14 schematisch dargestellt.

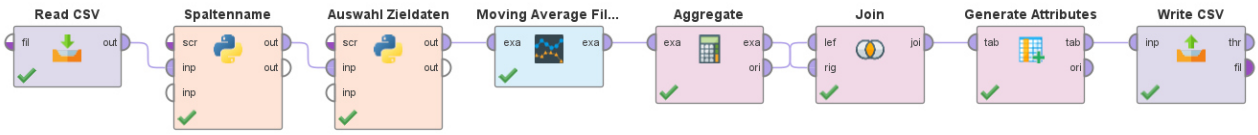

Abbildung 14: Ablauf in RapidMiner Studio

Abbildung 14 veranschaulicht die Aufbereitung der Rohdaten sowie die Umsetzung von Phase 2 und Phase 3 des Vorgehensmodells nach Fayyad et al. (1996) in RapidMiner Studio. Der Prozess beginnt mit dem Einlesen der Rohdaten, gefolgt von der Ergänzung der Spaltennamen mittels eines Python-Skripts. In Phase 2 erfolgt die Erstellung des Zieldatensatzes durch Identifikation und Entfernung von Sensoren, die nicht informativ sind. Hierzu analysiert ein weiteres Python-Skript die Varianz aller Sensoren und entfernt diejenigen, die einen nahezu konstanten Wert beibehalten. Als Schwellenwert für den Varianzfilter wurde $0{,}001$ empirisch bestimmt. Ein höherer Wert führte zur ungewollten Entfernung scheinbar relevanter Sensoren. Die Festlegung des Schwellenwerts basierte auf einer Analyse der Varianzverteilung über alle

Sensoren. Für Phase 3 zur Datenbereinigung und Datenvorverarbeitung wird Rauschentfernung durch den in RapidMiner Studio integrierten „Moving Average Filter“ Operator realisiert. Die gewählte Fensterbreite von drei Zyklen gewährleistet hier eine effektive Glättung bei gleichzeitiger Erhaltung des Trends der Degradation. Abschließend wird die Zielvariable RUL berechnet. Hierfür wird zunächst für jedes Triebwerk die maximale Zyklusanzahl in einer separaten Spalte erfasst. Die RUL-Variable folgt einem linearen Verlauf, beginnend bei der maximalen Zyklusanzahl des ersten aufgezeichneten Zyklus und nimmt schließlich konstant mit jedem vergangenen Zyklus ab bis zum Erreichen von null, wodurch der Ausfall des Triebwerks markiert ist. Für jedes nachfolgende Triebwerkt beginnt dieser Ablauf entsprechend seiner jeweiligen maximalen Zyklusanzahl erneut.

Der experimentelle Vergleichsrahmen ist für relationale Datenbanken ausgelegt, weswegen in einem nächsten Schritt die Erstellung der relationalen Datenbank erfolgt. Umgesetzt wird die Erstellung mit einem Python-Skript und den Bibliotheken Pandas (Version 2.0.3), SQLite3 (Version 2.6.0) und NumPy (Version 1.24.4). Zunächst wird eine relationale Datenbank namens „nasa_turbofan.db“ erstellt. Diese initialisiert die grundlegende Datenbankstruktur und erstellt eine Tabelle für die wahren RUL-Werte, die wiederum mit den Triebwerkkennzahlen verknüpft werden. Hierfür werden die bereits vorverarbeiteten Trainingsdaten aus der CSV-Datei und die Testdaten aus der originalen TXT-Datei eingelesen. Die entstandene relationale Datenbank kann schließlich für den folgenden Abschnitt genutzt werden, in welchem die Implementierung genauer erläutert wird.

5.2.2 Implementierung

Die technische Umsetzung des experimentellen Vergleichrahmens erfolgt in Python (Version 3.8.20) unter Verwendung der Bibliotheken scikit-learn (Version 1.6.0), Pandas (Version 2.0.3), NumPy (Version 1.24.4), Matplotlib (Version 3.10.3) und die Standardbibliotheken SQLite3, os, time und tracemalloc. Die Implementierung folgt einem modularen Ansatz und besteht aus den drei Komponentenarten des Metriken-Frameworks, den austauschbaren Vorverarbeitungsmodulen und dem Hauptskript. Zur Übersicht sind die unterschiedlichen Module hier gelistet und werden im Weiteren erläutert:

- Metriken-Framework: metrics.py
- Vorverarbeitungsmodule: preprocessing_dt.py, preprocessing_dr.py
- Hauptskripte: dr_iso_modul.py, dt_iso_modul.py, kombi_modul.py

Das Metriken-Framework implementiert die drei in dem Vergleichsrahmen definierten Messblöcke, sowie die zugehörigen Vergleichsblöcke. Unter Algorithmus 1 ist die Klasse MetricsStorage angegeben, durch die eine strukturierte Datenhaltung realisiert wird.

Algorithmus 1: Ausschnitt zur Klasse MetricsStorage

```
1. class MetricsStorage:
2.     def __init__(self):
3.         self.rohdaten    = {}
4.         self.verfahren_1 = {}
5.         self.verfahren_2 = {}
6.         self.vergleich   = {}
7.     def get_verfahren_storage(self, verfahren_name):
8.         if verfahren_name    == „verfahren_1“:
9.             return self.verfahren_1
10.            elif verfahren_name  == „verfahren_2“:
11.            return self.verfahren 2
```

In Algorithmus 1 ist zu sehen, dass die Klasse MetricsStorage sowohl die erfassten Charakteristika der Rohdaten als auch die Messergebnisse der vorverarbeiteten Daten für spätere Berechnungen und Vergleiche verfügbar hält. Die Messungen zur Charakterisierung der Rohdatenbasis werden in dem ersten Messblock durchgeführt. Unteranderem werden hier die Korrelationsstruktur zur Erfassung der Interdependenzen zwischen den Sensorsignalen sowie die Verteilungseigenschaften, die sich mithilfe des Schiefemaßes und des Shapiro-Wilk-Test prüfen lassen, aufgenommen. Letzterer ist für Kapazitätsoptimierung auf maximal 5000 Datenpunkte beschränkt. Der zweite Messblock führt die gleichen Messungen auf die vorverarbeiteten Daten durch und erweitert diese zusätzlich um die notwendigen Messungen für Vergleichsblock B. Die Vergleichsblöcke implementieren die eigentliche Auswertungslogik und setzen die Prüfmethode um.

Die Vorverarbeitungsmodule implementieren die ausgewählten Vorverarbeitungsverfahren und erlauben über ein Slot-System in den jeweiligen Hauptskripten den Austausch verschiedener Verfahren. Jedes Verfahren wird über den einheitlichen Methodenaufruf „vorverarbeitungs_slot“ angesprochen. In Algorithmus 2 ist hierzu ein Ausschnitt zu sehen, wodurch gerade der flexible Austausch verschiedener Verfahren ohne Codeänderungen ermöglicht wird.

Der Ausschnitt der Python-Funktion in Algorithmus 2 für die Datentransformation nimmt eine Eingabematrix X sowie einen Methodennamen als Parameter entgegen. Je nach gewähltem Verfahren wird dieses entsprechend aus der Bibliothek scikit-learn ausgewählt und verwendet. Anschließend wird das Ergebnis als neuer Data Frame mit den ursprünglich Spaltennamen und Indizes zurückgegeben. Dimensionsreduktionsverfahren erzeugen neue Merkmale, weswegen das Vorverarbeitungsmodul für die Datenreduktion zusätzlich eine einheitliche Behandlung der hierbei resultierenden Merkmalsnamen berücksichtigt.

Algorithmus 2: Ausschnitt Slot-Funktion für die Vorverarbeitung

```
1. def vorverarbeitungs_slot(X, methode_name="MinMax"):
2.     print(f"\n VORVERARBEITUNG: {methode_name}")
3.
4.     if methode_name == "MinMax":
5.         transformer = MinMaxScaler()
6.    ...
7.     else:
8.         print(f"Unbekannte Methode: {methode_name}")
9.         return X
10.
11.    X_transformed = pd.DataFrame(
12.           transformer.fit_transform(X),
13.           columns=X.columns,  # Spaltennamen beibehalten
14.           index=X.index       # Index beibehalten
15.       )
```

Die Hauptskripte organisieren die Ausführung der verschiedenen Vorverarbeitungsverfahren und koordinieren die Erfassung der Messungen. Für den isolierten Vergleich existieren die separaten Skripte für Datentransformation und Datenreduktion, die jeweils die gesamte Pipeline von dem Einlesen der Rohdaten über die spezifischen Vorverarbeitungsverfahren bis hin zum Data-Mining-Algorithmus und der jeweils anschließenden Evaluation implementieren. Der kombinierte Vergleich wird durch ein integriertes Skript realisiert, das beide Vorverarbeitungsschritte in verschiedenen Reihenfolgen ausführt und ebenfalls die Zwischenergebnisse für die Evaluation erfasst. Die Ausgabe der Vergleichsergebnisse erfolgt in tabellarischer Form innerhalb des Terminals. Des Weiteren wird eine relationale Datenbank als Eingabe verwendet. Als Beispiel ist ein Ausschnitt des Hauptskriptes für die Datentransformation in Algorithmus 3 zu sehen.

Der in Algorithmus 3 zu sehende Codeabschnitt konstruiert zunächst den Pfad zur Datenbankdatei. Anschließend wird eine Verbindung zur SQLite-Datenbank hergestellt und drei separate SQL-Abfragen ausgeführt, um die Trainings-, Testdaten und die wahren RUL-Werte in entsprechende Pandas Data Frames zu laden. Nach erfolgreichem Laden aller Daten wird die Datenbankverbindung wieder geschlossen. Nachdem die Implementierung abgeschlossen ist, wird im folgenden Abschnitt die Umsetzung des experimentellen Vergleichrahmens durchgegangen.

Algorithmus 3: Ausschnitt aus dem Skript für den isolierten Datentransformationsvergleich

```
1. # Laden der Daten
2. db_path = os.path.join(os.path.dirname(__file__), 'nasa_turbo-
   fan.db')
3. conn = sqlite3.connect(db_path)
4.
5. try:
6.     train_df = pd.read_sql_query("SELECT * FROM train_data",
       conn)
7.
8.     test_df = pd.read_sql_query("SELECT * FROM test_data",
       conn)
9.
10.    true_rul_df = pd.read_sql_query("SELECT * FROM true_rul",
       conn)
11.   ...
12.   conn.close()
```

Im folgenden Kapitel wird auf die Anwendung der Datentransformations- und Datenreduktionsverfahren eingegangen.

5.2.3 Anwendung der Datentransformations- und Datenreduktionsverfahren

Die Anwendung der Datentransformations- und Datenreduktionsverfahren wird durch den experimentellen Vergleichsrahmen durchlaufen (vgl. Abschnitt 4.4). Darüber hinaus entspricht dieses Kapitel gerade Phase 4 des Vorgehensmodells nach Fayyad et al. (1996). Die Rohdaten, die in Abschnitt 5.2.1 in eine relationale Datenbank überführt wurden, werden in einem ersten Schritt eingelesen. Hiernach folgt Messblock 1, der zur Charakterisierung der Rohdaten dient und die erste Messgrundlage für den Vergleich durchführt.

Anschließend folgt der Vorverarbeitungsblock. Hierin kommen die jeweiligen Vorverarbeitungsverfahren gemäß dem gewählten Versuchsszenario zum Einsatz. Es werden insgesamt drei Konfigurationen untersucht. Zum einen die isolierte Anwendung von Datentransformationsverfahren, weiter noch die isolierte Anwendung von Datenreduktionsverfahren sowie deren Kombination. Für die Datentransformation werden die Z-Wert-Normalisierung und die Min-Max-Normalisierung verwendet, während bei der Datenreduktion PCA und FastICA zum Einsatz kommen (vgl. Abschnitt 5.1). Im kombinierten Vergleich werden diejenigen Verfahren kombiniert, die sich im isolierten Vergleich als besonders leistungsfähig erwiesen haben.

Nach den Verfahrensblöcken geht es in den Messblock 2 über. Hier erfolgen die Messungen für die Metriken von Vergleichsblock A und Vergleichsblock B. Um die Darstellung des Experiments übersichtlich zu halten und inhaltliche Brüche zu vermeiden, werden die Ergebnisse

jeweils direkt im Anschluss an ihre Erhebung präsentiert. Die Abbildung 15 zeigt den Vergleich der Datentransformationsverfahren anhand der entwickelten Metriken.

Kriterium	Min-Max	Z-Wert
Korrelationserhaltung	1.000	1.000
Informationserhaltung	1.000	0.999
Skalierungsheterogenität	0.000	0.184
Merkmalsinterdependenz	0.583	0.583
Streuungsnutzungsgrad	0.133	0.133
Numerische Stabilität	9.63e+01	1.75e+01
Varianzerhaltung	0.866	0.873
Cluster-Trennbarkeit	0.129	0.156

Abbildung 15: Vergleich der Datentransformationsverfahren anhand der entwickelten Metriken

In Abbildung 15 ist zu sehen, dass die Metriken zur Normalitäts- und Symmetrieverbesserung entfallen sind. Diese Metriken gehen explizit auf die Verteilungsoptimierung ein (vgl. Abschnitt 4.3). Da nur lineare Datentransformationen betrachtet werden, findet kein Einfluss auf die Verteilung der Daten statt. Der Vollständigkeit halber wurden diese Metriken dennoch im Code implementiert. Die Ergebnisse des Datentransformationsvergleichs zeigen, dass beide Verfahren die Korrelationserhaltung vollständig bewahren. Die Informationserhaltung liegt bei beiden Verfahren nahezu bei optimalen Werten. Bei der Skalierungsheterogenität zeigt die Min-Max-Normalisierung eine perfekte Homogenisierung, während die Z-Wert-Normalisierung einen moderaten Wert von $0{,}184$ aufweist. Die Merkmalsinterdependenz und der Streuungsnutzungsgrad bleiben bei beiden Verfahren identisch. Die numerische Stabilität unterscheidet sich hingegen deutlich, wobei die Z-Wert-Normalisierung eine bessere Konditionierung aufweist als es die Min-Max-Normalisierung tut. Die Varianzerhaltung liegt bei beiden Verfahren in ähnlichen Bereichen. Ebenso die Cluster-Trennbarkeit. In Abbildung 16 ist der Vergleich der Datenreduktionsverfahren dargestellt.

Kriterium	PCA	FastICA
Informationserhaltung	0.245	0.226
Varianzerhaltung	1.000	0.715
Numerische Stabilität	8.40e+01	1.00e+00
Cluster-Trennbarkeit	0.269	0.109
Skalierungsheterogenität	1.665	0.085
Merkmalsinterdependenz	0.000	0.000
Streuungsnutzungsgrad	0.122	0.122

Abbildung 16: Vergleich der Datenreduktionsverfahren anhand der entwickelten Metriken

Der Vergleich in Abbildung 16 zeigt deutliche Unterschiede zwischen PCA und FastICA. Die Informationserhaltung liegt bei beiden Verfahren auf niedrigem Niveau. Die Varianzerhaltung unterscheidet sich wiederum erheblich. Hierbei hat die PCA eine vollständige Erhaltung erreicht und FastICA nur $0{,}715$. Bei der numerischen Stabilität zeigt FastICA eine bessere Konditionierung als die PCA. Die Cluster-Trennbarkeit ist bei PCA höher als bei FastICA. Auffällig ist die Skalierungsheterogenität, die bei der PCA einen hohen Wert von $1{,}665$ zeigt. Die FastICA schneidet hierin mit $0{,}085$ deutlich besser ab. Beide Verfahren eliminieren die Merkmalsinterdependenz vollständig und zeigen identische Werte beim Streuungsnutzungsgrad. In der folgenden Abbildung 17 ist der kombinierte Vergleich der Z-Wert-Normalisierung und PCA zu sehen.

Kriterium	Kombi. 1	Kombi. 2
Informationserhaltung	0.384	0.244
Varianzerhaltung	0.954	0.715
Numerische Stabilität	5.57e+00	1.00e+00
Cluster-Trennbarkeit	0.203	0.109
Skalierungsheterogenität	0.697	0.117
Merkmalsinterdependenz	0.000	0.000
Streuungsnutzungsgrad	0.128	0.122

Abbildung 17: Kombinierter Vergleich anhand der entwickelten Metriken

Der kombinierte Vergleich in Abbildung 17 zeigt die Auswirkungen der Reihenfolge der Verfahren. In Kombination 1 wird zunächst die Z-Wert-Normalisierung durchgeführt, gefolgt von der Anwendung der PCA. Für Kombination 2 ist die umgekehrte Reihenfolge festgelegt worden. Die Informationserhaltung ist bei Kombination 1 mit $0{,}384$ deutlich höher als bei Kombi-

nation 2 mit $0{,}244$. Die Varianzerhaltung ist bei Kombination 1 etwas erhöhter als bei Kombination 2. Die numerische Stabilität ist bei Kombination 1 mit schlechter als bei Kombination 2. Die Cluster-Trennbarkeit liegt bei Kombination 1 höher als bei Kombination 2. Bei der Skalierungsheterogenität zeigt Kombination 1 einen höheren Wert von $0{,}697$ im Vergleich zu Kombination 2 mit $0{,}117$. Beide Kombinationen eliminieren die Merkmalsinterdependenzen vollständig und weisen bei dem Streuungsnutzungsgrad ähnliche Werte auf.

5.2.4 Modellierung

In diesem Kapitel werden die Phasen 5 bis 7 des Vorgehensmodells nach Fayyad et al. (1996) durchlaufen. Da beabsichtigt ist den Datensatz für prädiktive Instandhaltung zu nutzen wird und das Analyseziel die Vorhersage der verbleibenden Nutzungsdauer ist, wurde ein Regressionsverfahren gewählt. Zum Einsatz kommt ein SVM-Regressor, der für kontinuierliche Zielgrößen geeignet ist und sich in der Literatur für prädiktive Instandhaltungsaufgaben bewährt hat (vgl. Abschnitt 2.4). Abweichend vom Vorgehensmodell nach Fayyad et al. (1996) wird in dieser Arbeit bewusst nur ein Data-Mining-Algorithmus eingesetzt. Ziel der Arbeit ist nicht der Vergleich verschiedener Modellierungsverfahren, sondern der Vergleich von Verfahren der Datentransformation und Datenreduktion. Die, wie in dem experimentellen Vergleichsrahmen vorgesehene Festlegung auf einen Algorithmus ermöglicht eine kontrollierte Bewertung der Vorverarbeitungsschritte. Im experimentellen Vergleichsrahmen entspricht Phase 7 des Vorgehensmodells nach Fayyad et al. (1996) gerade der Umsetzung des Data-Mining-Blocks, aus dem jeweils zwei Modelle pro Vergleichsszenario resultieren. In Messblock 3 werden die Modelle abschließend evaluiert. Zur Visualisierung der Modellgüte werden für jedes Versuchsszenario zwei Scatterplots dargestellt. Diese zeigen auf der x-Achse die tatsächlichen RUL-Werte und auf der y-Achse die durch das Modell vorhergesagten Werte. In Abbildung 18 ist der Vergleich der Datentransformationsverfahren hinsichtlich des entstandenen Data-Mining-Models zu sehen.

Die Ergebnisse der SVM-Regression zeigen in Abbildung 18 klare Leistungsunterschiede. SVM erreicht mit der Z-Wert-Normalisierung einen MAE von 25,7 bei einem R^2 von 0,366, während SVM mit der Min-Max-Normalisierung eine bessere Performanz mit einem MAE von 19,3 und einem R^2 von 0,624 aufweist. Die Streudiagramme verdeutlichen die unterschiedliche Vorhersagequalität, wobei die Min-Max-Normalisierung eine stärkere lineare Beziehung zwischen den wahren und vorhergesagten Werten zeigt.

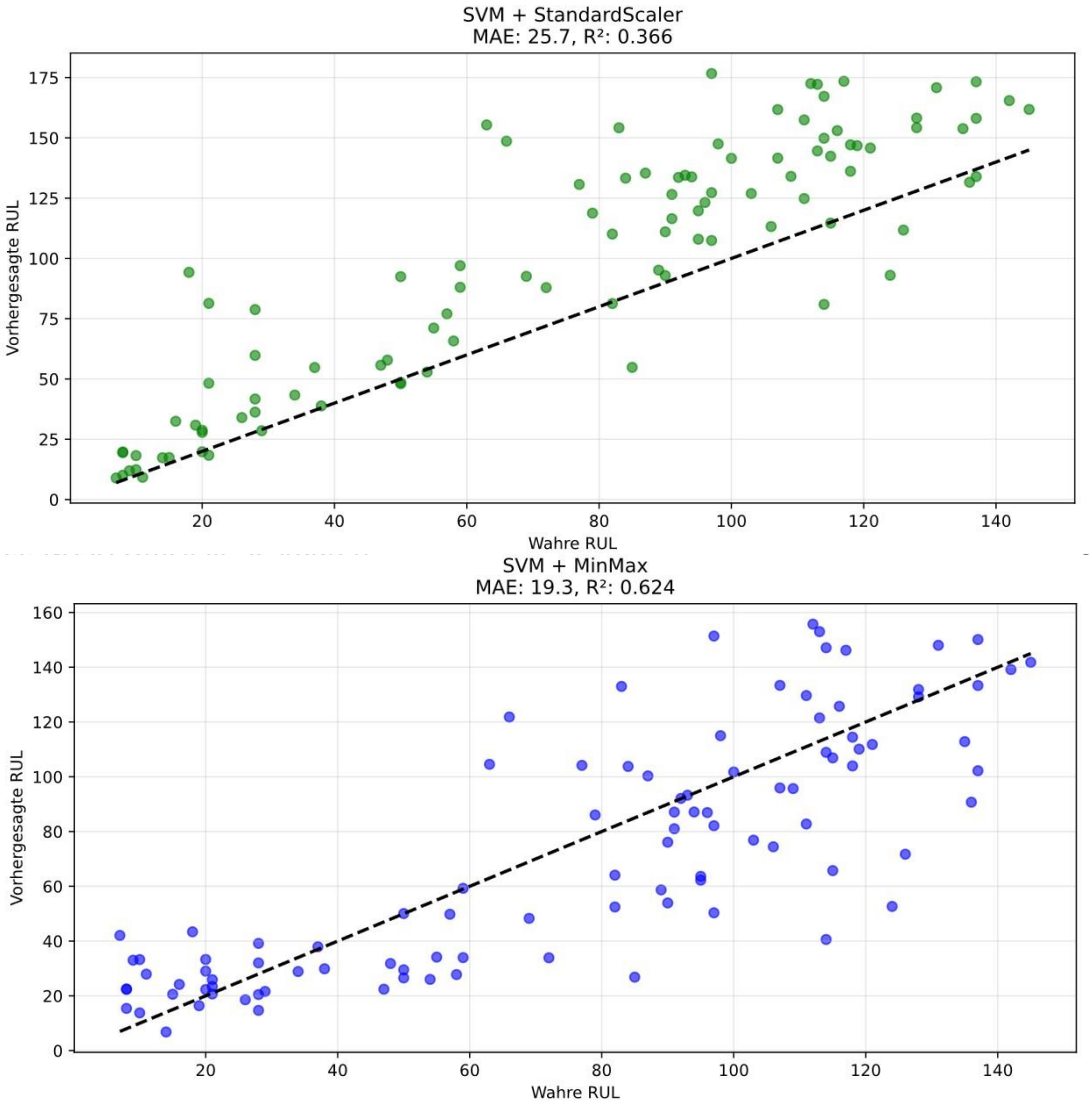

Abbildung 18: Streudiagramme zur Vorverarbeitung mit Datentransformation

Die nachfolgende Abbildung 19 zeigt die Ergebnisse der SVM-Regression nach Anwendung der zu vergleichenden Datenreduktionsverfahren.

Abbildung 19 zeigt insgesamt, dass die SVM-Regression bei der Verwendung von Datenreduktionsverfahren deutlich schlechter abschneidet als bei der alleinigen Anwendung von Datentransformationsverfahren. Unter Verwendung der PCA wird ein MAE von $59{,}4$ bei einem stark negativen R^2 von $-2{,}335$ erreicht. Dies deutet auf eine sehr schlechte Modellanpassung hin. SVM mit FastICA zeigt mit einem MAE von $47{,}2$ und einem R^2 von $-0{,}935$ zwar bessere, jedoch noch immer unzureichende Ergebnisse. Beide Ansätze führen zu einer wesentlich schlechteren Vorhersagequalität.

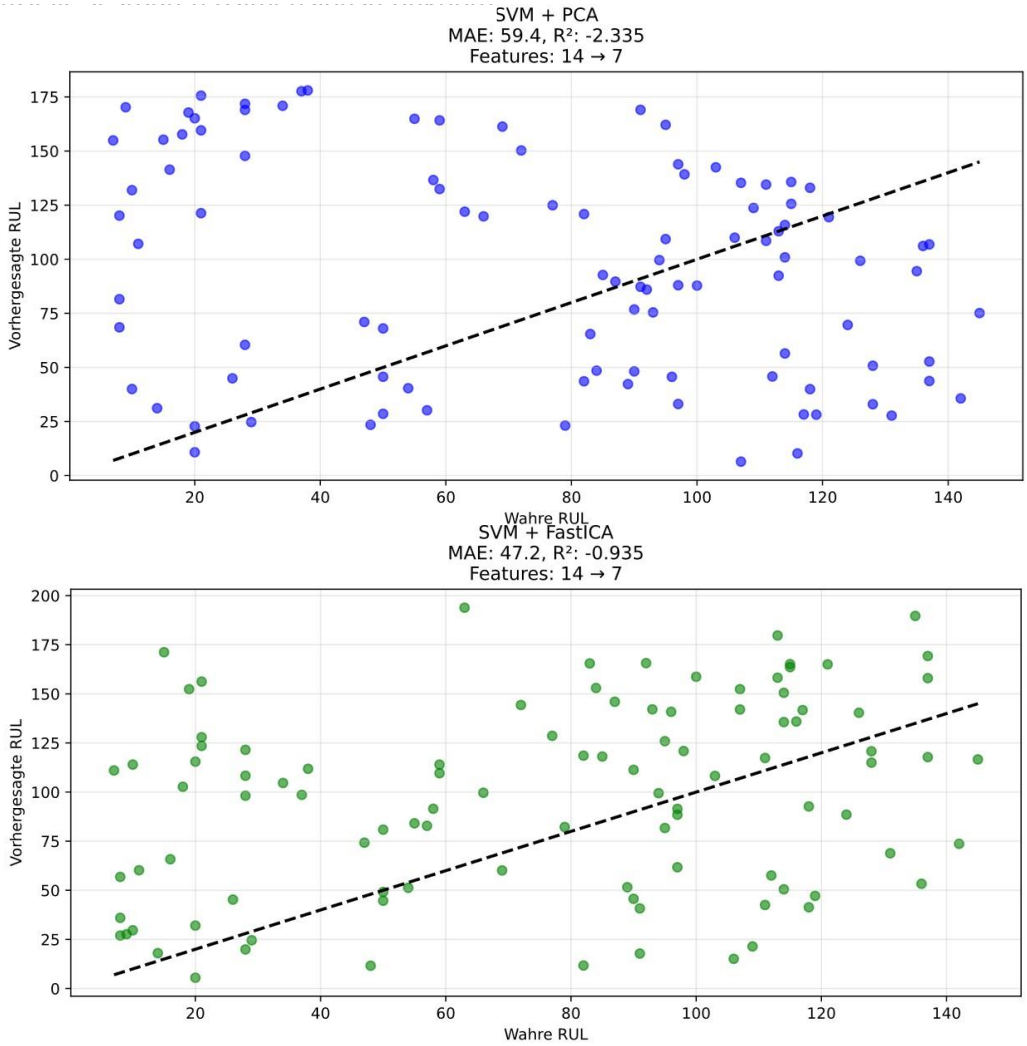

Abbildung 19: Streudiagramme zur Vorverarbeitung mit Datenreduktion

In Abbildung 20 werden schließlich die kombinierten Ansätze von Datentransformation und Datenreduktion dargestellt.

Die kombinierten Verfahren zeigen in Abbildung 20 ebenfalls deutliche Unterschiede hinsichtlich ihrer Verfahrensreihenfolge. Die SVM mit Z-Wert-Normalisierung vor der anschließenden PCA erziel einen MAE von $28{,}0$ und einen R^2 von $0{,}274$. Wird die Reihenfolge umgekehrt, so verschlechtert sich die Modelleistung deutlich. Der MAE steigt auf $58{,}7$ und der Wert für R^2 sinkt auf $-2{,}189$.

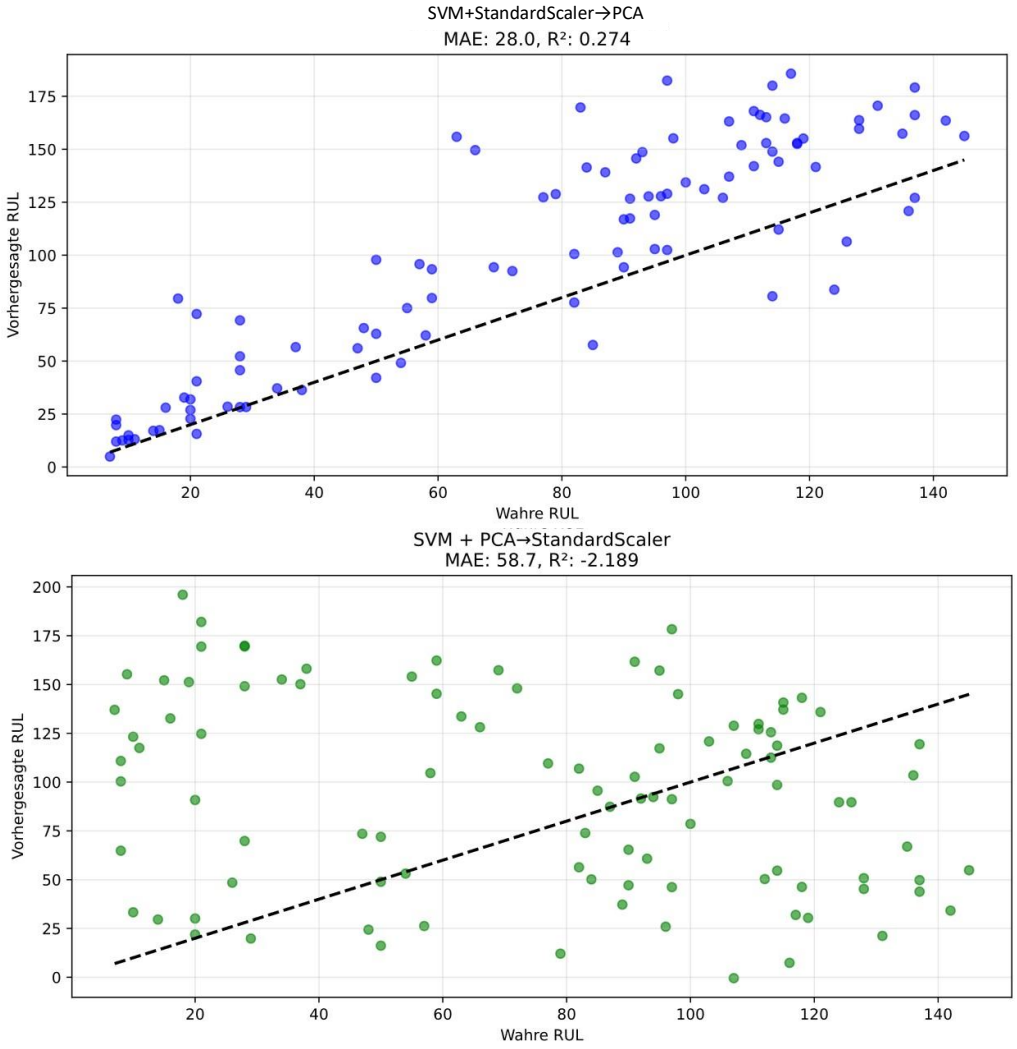

Abbildung 20: Streudiagramme zur Vorverarbeitung mit Datentransformation und Datenreduktion

5.2.5 Ergebnisanalyse

Im Bereich der Datentransformation zeigt sich eine klare Übereinstimmung zwischen den Metriken und der Modelleistung. Die Min-Max-Normalisierung führt durch ihre vollständige Skalierungshomogenität zu den besten Ergebnissen bei der SVM. Weiter noch erzielt die Z-Wert-Normalisierung zwar bessere Werte hinsichtlich numerischer Stabilität und Cluster-Trennbarkeit, führt jedoch in der Praxis zu einer schwächeren Modelleistung. Hieraus lässt sich schließen, dass nicht alle in den Metriken vorteilhaften Eigenschaften den gleichen Einfluss auf das praktische Ergebnis haben. Im Falle der Datenreduktion ist der starke Informationsverlust entscheidend für die deutlich schlechteren Regressionsleistungen. Beide getesteten Datenreduktionsverfahren weisen negative Werte für R^2 auf und zudem zeigen sie einen erheblichen Anstieg des MAE. Interessanterweise schneidet FastICA trotz geringer Varianzerhaltung besser ab als die PCA. Dies deutet darauf hin, dass statistische Unabhängigkeit der Komponenten für bestimmte Aufgaben relevanter ist als eine maximale Varianzerklärung.

Bei der kombinierten Anwendung von Vorverarbeitungsverfahren zeigt sich, dass die Reihenfolge der Schritte einen großen Unterschied macht. Wenn zuerst die Datentransformation und anschließend die Datenreduktion durchgeführt wird, so sind die Ergebnisse sowohl in den Metriken als auch in der Modellgüte deutlich besser. Die umgekehrte Reihenfolge führt zu einer erheblichen Verschlechterung. Dies zeigt, dass eine anhand der Metriken gestützte Auswahl und Reihenfolge der Vorverarbeitungsschritte zum Erfolg des Data-Mining-Schrittes beiträgt. Die Analyse deutet darauf hin, dass einige Metriken einen starken Einfluss auf die Bewertung der Verfahren haben. Hierzu erweist sich die Skalierungsheterogenität als besonders geeignet, um den Erfolg bei distanzbasierten Verfahren vorherzusagen. Die Informationserhaltung ist ein verlässlicher Indikator für die Eignung von Reduktionsverfahren. Zugleich wird ersichtlich, dass sich nicht alle Metriken unmittelbar in einer verbesserten Modellleistung widerspiegeln. Beispielsweise hat die numerische Stabilität keinen klaren Einfluss auf die Modellleistung gezeigt.

Aus den Erkenntnissen ergeben sich abschließend konkrete Handlungsempfehlungen für das Fallbeispiel. Bei distanzbasierten Modellen sollte besonderer Wert auf eine gleichmäßige Normalisierung gelegt werden. Wenn eine Reduktion angewendet werden soll, sollte FastICA bevorzugt werden, da es sich als robuster gegenüber Informationsverlust erwiesen hat. Werden Datentransformation und Datenreduktion kombiniert, so sollte stets die Datentransformation an erster Stelle stehen. Insgesamt zeigt sich, dass der experimentelle Vergleichsrahmen, mit den Metriken und der praktischen Validierung im Data-Mining-Schritt, dabei unterstützen kann, Vorverarbeitungsverfahren besser zu verstehen und fundierte Entscheidungen im Wissensentdeckungsprozess zu treffen.

5.3 Diskussion und Fazit

Die Entwicklung des vorgestellten Vergleichrahmens beabsichtigt eine Lücke im systematischen Vergleich und der Bewertung von Datenvorverarbeitungsverfahren der Datentransformation und Datenreduktion im Kontext von Data Mining zu schließen. Während bestehende Studien entweder die Datenreduktion oder ausgewählte Aufgaben der Datentransformation separat untersuchen, bleibt insbesondere unklar, wie sich diese Schritte gegenseitig beeinflussen und unter welchen Bedingungen sie sinnvoll kombiniert werden können. Im Rahmen dieser Arbeit wurden zahlreiche Randbedingungen identifiziert, die für eine Bewertung der Datentransformation und Datenreduktion berücksichtigt werden müssen. Besonders relevant sind hierbei die Herausforderungen der Diskrepanz zwischen den Rohdaten und algorithmischen Anforderungen sowie der Effekt unterschiedlicher Reihenfolgen bei der Anwendung der Datentransformations- und Datenreduktionsverfahren. Eine große Rolle bei der Entwicklung des Vergleichsrahmens hat die Entwicklung und Auswahl geeigneter, aussagekräftiger Metriken eingenommen. Diese machen die Wirksamkeit der Verfahren präzise und vergleichbar

messbar. Der entwickelte Vergleichsrahmen zur Bewertung der Verfahren ist für den Wissensentdeckungsprozesses in dem Vorgehensmodell nach Fayyad et al. (1996) verankert. Eine weitere Herausforderung ist die Auswahl vergleichbarer Aufgaben der Datentransformation. Hierzu wurde zunächst die Einschränkung zur Betrachtung verfahrensspezifischer Transformationen, die auf Datenebene ablaufen, getroffen. Weiter konnte festgestellt werden, dass insbesondere lineare Transformationen und Verteilungstransformationen berücksichtigt werden müssen, da innerhalb dieser Aufgaben unterschiedliche Verfahren in der Literatur zu finden sind, die differenzierbare Ergebnisse liefern, sodass ein Vergleich unter diesen möglich ist. Im Bereich der Datenreduktion sind die Verfahren weitgehend vergleichbar, da sie dasselbe Ziel einer Reduktion der Datenmenge bei möglichst geringem Informationsverlust verfolgen.

Der Vergleichsrahmen wurde in einem mehrstufigen Prozess entwickelt, der auf einer kritischen Gegenüberstellung und Operationalisierung von Randbedingungen, Anforderungen der betrachteten Datenvorverarbeitungsverfahren und Anforderungen des Data Mining aufbaut. Dies gewährleistet ein strukturiertes und fundiertes Vorgehen, wodurch die tatsächlich vergleichbaren Metriken erfolgreich identifiziert werden konnten. Der in Abschnitt 4.2 entwickelte Kriterienkatalog bietet eine strukturierte Herangehensweise zur Identifikation relevanter Anforderungen der Datenvorverarbeitung und des Data Mining. In dieser Arbeit wurden die so bestimmten Anforderungen genutzt, um Rückschlüsse auf Metriken für den Erfolg der Datenvorverarbeitung zu ziehen. Hierzu ist eine konzeptuelle Umkehrung der Anforderungslogik vorgenommen worden. Mit Hilfe der Anforderungen konnte folglich das entscheidende Vorverarbeitungsziel und entsprechend die Operationalisierung der Anforderungen aufgestellt werden. Diese Untersuchung führt zu einer Gliederung in verschiedene Dimensionen, anhand derer die Metriken zur sinnvollen Messung der jeweiligen Dimension strukturiert hergeleitet werden können. Die so entstandenen Metriken wurden schließlich in übergeordnete Kategorien eingegliedert, die eine Zuordnung zu isolierten und kombinierten Vorgehensweisen ermöglichen. Nur auf diese Weise konnte die Einteilung in drei Messblöcke und zugehörige Vergleichsblöcke erfolgen. Diese Vorgehensweise ermöglicht die Bewertung an verschiedenen Punkten des Wissensentdeckungsprozesses. Dies wiederum gewährleistet eine ganzheitliche Perspektive auf die Auswirkungen der Vorverarbeitung. Darüber hinaus wurde auf diese Art und Weise festgestellt, dass für den kombinierten Vergleich die Metriken zur Kategorie der Verteilungsoptimierung keine weitere Aussagekraft haben.

Des Weiteren ist für den Vergleichsrahmen eine konstante experimentelle Umgebung festgelegt worden. Durch die Beibehaltung des Rohdatensatzes und eines Data-Mining-Algorithmus über alle Experimente hinweg, wurde sichergestellt, dass die gemessenen Effekte tatsächlich auf die Datenvorverarbeitung zurückzuführen sind und kein Maßstabsvergleich des Data-Mi-

ning-Algorithmus selbst stattfindet. Hierdurch wird eine interne Validität der Ergebnisse gewährleistet. Insgesamt hat die Arbeit eine Vorgehensweise etabliert, die es ermöglicht die Auswirkung der Datentransformation und Datenreduktion objektiv zu bewerten. Die konsequente Ableitung von Zielen und Herleitung von Metriken sorgt für eine entsprechend strukturierte Herangehensweise.

Eine inhärente Limitation des entwickelten Vergleichsrahmens liegt in der eingeschränkten Generalisierbarkeit bestimmter Datentransformationsverfahren. Die Arbeit schließt bewusst stark domänenspezifische oder deterministische Datentransformationen aus. Auch Aufgaben der Datentransformation, die eine semantische und keine strukturelle Anpassung der Daten vornehmen, werden ausgenommen. Dies ist eine methodisch notwendige Entscheidung, um die Komplexität des Vergleichs zu handhaben sowie objektive und messbare Unterscheidungen erkennbar zu machen. Ein Vergleich und die Bewertung der ausgeschlossenen Verfahren würde in vielen Fällen Expertenwissen erfordern und die Entwicklung allgemeingültiger Metriken signifikant erschweren.

Die Anwendbarkeit des Vergleichsrahmens mit den zugehörigen Metriken zur Bewertung der untersuchten Vorverarbeitungsverfahren wurde durch die Anwendung in Abschnitt 5.2 gezeigt. Diese wird anhand eines realitätsnahen Datensatzes durchgeführt. Hierin wird deutlich, dass die Einschätzung basierend auf dem Vergleichsrahmen auch der realen Umsetzbarkeit entspricht. Die modulare Struktur der Implementierung stellt sicher, dass eine flexible Erweiterung möglich ist. Neue Verfahren können einfach hinzugefügt und bestehende ausgetauscht werden, wodurch die langfristige Nutzbarkeit des implementierten Vergleichsrahmens sichergestellt werden soll.

Das Fazit dieser Arbeit ist ein neu entwickelter experimenteller Vergleichsrahmen, der den systematischen und quantitativen Vergleich von Datentransformations- und Datenreduktionsverfahren im Kontext von Data Mining ermöglicht. Durch die Anwendung des Vergleichsrahmens konnte die konkrete Auswirkung der Abfolge von Datentransformation und Datenreduktion auf die Leistungsfähigkeit nachgelagerter Data-Mining-Algorithmen umfassend bewertet und deren Rolle für valide und belastbare Analyseergebnisse im Wissensentdeckungsprozess belegt werden. Die Anwendung des experimentellen Vergleichrahmens bestätigt die praktische Relevanz der Metriken für die Bewertung von Datenvorverarbeitungsverfahren. Die Vorhersagen der Metriken korrelierten mit den Ergebnissen des Data Mining. Besonders zeigte sich dies bei der Skalierungsheterogenität, bei der die perfekte Homogenisierung der Min-Max-Normalisierung direkt mit der Regressionsleistung korrespondierte. Gleichermaßen spiegelten sich die niedrigen Werte der Informationserhaltung der Datenreduktionsverfahren unmittelbar in einer schlechteren Modelleistung mit negativen R^2-Werten wider. Es konnte belegt werden,

dass die Reihenfolge und Kombination von Datentransformation und Datenreduktion entscheidend auf die Datenqualität und Modelleistung wirkt.

6 Zusammenfassung und Ausblick

Die vorliegende Arbeit entwickelt einen experimentellen Vergleichsrahmen für Datentransformations- und Datenreduktionsverfahren im Data-Mining-Kontext. Dieser wird basierend auf dem Vorgehensmodell nach Fayyad et al. (1996) umgesetzt. Zur Operationalisierung der unterschiedlichen Anforderungen während des gesamten Prozesses der Datenvorverarbeitung, entstand zunächst ein dreistufiger Kriterienkatalog. Die Einhaltungskriterien dienten als binärer Filter, um nicht anwendbare Verfahren frühzeitig auszuschließen. Weiter repräsentierten die Einflusskriterien kontextabhängige Faktoren, die die Wirksamkeit der Verfahren unter spezifischen Ausgangsdaten einordnet. Abschließend sind die Evaluationskriterien erläutert worden. Sie dienen der Einschätzung dazu wie der Erfolg der Verfahren zu quantifizieren ist. Die Metriken wurden mithilfe des Kriterienkatalogs entwickelt und anschließend übergeordneten Kategorien zugeordnet. Diese Metriken wurden dann in einem Vergleichsrahmen mit definierten Mess- und Vergleichsblöcken eingebettet. Der Vergleichsrahmen erlaubt sowohl die isolierte Betrachtung einzelner Vorverarbeitungsschritte als auch die kombinierte Anwendung von Datentransformation und Datenreduktion. Für die Datentransformation wurden Verfahren der linearen Transformation und Verteilungstransformationen ausgewählt, sodass quantifizierbare Unterschiede gemessen werden konnten.

Zur Validierung wurde der Vergleichsrahmen in Python implementiert und exemplarisch auf den NASA Turbofan Engine Degradation Datensatz angewendet. Diese Fallstudie erlaubt eine praktische Erprobung unter kontrollierten Bedingungen. Es wurde ein gleichbleibender Rohdatensatz und ein konstanter Data-Mining-Algorithmus berücksichtigt. Weiter sind neben der isolierten Durchführung auch Verfahrensketten analysiert worden, um Reihenfolgeeffekte und Synergien aufzudecken. Basierend auf den Erkenntnissen aus der Entwicklung des Vergleichsrahmens und aus dem Fallbeispiel ist abschließend eine Diskussion geführt worden.

Der entwickelte Vergleichsrahmen legt den Grundstein für fortführende Forschungsansätze. Die Ausweitung des Anwendungskontextes könnte weiterführende Erkenntnisse liefern. Bislang wurde der Vergleich auf einen spezifischen Data-Mining-Algorithmus angewandt, um die interne Validität zu gewährleisten. Die Ausweitung auf unterschiedliche Data-Mining-Algorithmen würden die Generalisierbarkeit der Ergebnisse erhöhen und könnte insbesondere dazu beitragen differenziertere Handlungsanweisungen zu treffen. In gleicher Weise könnte die Erprobung auf unterschiedlichen Datensätzen aus verschiedenen Domänen die Robustheit des Vergleichsrahmens prüfen.

Weiter noch könnte die Automatisierung der Prozesskette, also die Datenvorverarbeitungsverfahren mit anschließendem Data-Mining-Schritt, die Effizienz datenanalytischer Prozesse steigern. Aufbauend auf den Vergleichsergebnissen ließen sich Verfahren entwickeln, die eine

automatisierte Auswahl und Reihenfolge der Vorverarbeitungsschritte für gegebene Daten und Analyseziele ermöglichen.

Ein konkreter weiterführende Ansatz besteht zudem darin, mehr Verfahren im kombinierten Vergleich zu untersuchen. In Kapitel 5 wurde für die Kombination gezielt diejenigen Verfahren gewählt, die im isolierten Vergleich am besten abgeschnitten haben. Weitere Untersuchungen hierzu könnten zeigen, ob vermeintlich schwächere Verfahren in Kombination dennoch zu einer deutlichen Verbesserung der Ergebnisse führen als erwartet. Dies würde zusätzliche, neue Erkenntnisse über die Wechselwirkungen zwischen unterschiedlichen Vorverarbeitungsverfahren liefern.

7 Literaturverzeichnis

Abdallah, Zahraa S.; Du, Lan; Webb, Geoffrey I. (2020): Data Preparation. In: Dinh Phung, Geoffrey I. Webb und Claude Sammut (Hg.): Encyclopedia of Machine Learning and Data Science. New York, NY: Springer US, S. 1–10.

Ackoff, R. L. (1989): From Data to Wisdom. In: *Journal of Applied System Analysis* 16, S. 4–9.

Aggarwal, Charu C. (2014): Data Classification: Chapman and Hall/CRC.

Aggarwal, Charu C. (2017): Outlier Analysis. Cham: Springer International Publishing.

Ahsan, Md; Mahmud, M.; Saha, Pritom; Gupta, Kishor; Siddique, Zahed (2021): Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. In: *Technologies* 9 (3), S. 52. DOI: 10.3390/technologies9030052.

AlKarawi, Ahmed; AlJanabi, Kadhim (2022): Data Reduction Techniques: A Comparative Study. In: *Jour. Kufa Math. Comp.* 9 (2), S. 1–17. DOI: 10.31642/JoKMC/2018/090201.

Atkinson, Anthony C.; Riani, Marco; Corbellini, Aldo (2021): The Box–Cox Transformation: Review and Extensions. In: *Statist. Sci.* 36 (2). DOI: 10.1214/20-STS778.

Ayeche, Farid; Alti, Adel (2023): Efficient Feature Selection in High Dimensional Data Based on Enhanced Binary Chimp Optimization Algorithms and Machine Learning. In: *Human-Centric Intelligent Systems* 3 (4), S. 558–587. DOI: 10.1007/s44230-023-00048-w.

Bahri, Maroua; Bifet, Albert; Maniu, Silviu; Gomes, Heitor Murilo (2020): Survey on Feature Transformation Techniques for Data Streams. In: Marie desJardins und Christian Bessiere (Hg.): Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20}. Yokohama, Japan, 11.07.2020 - 17.07.2020. California: International Joint Conferences on Artificial Intelligence Organization, S. 4796–4802.

Bartenhagen, Christoph; Klein, Hans-Ulrich; Ruckert, Christian; Jiang, Xiaoyi; Dugas, Martin (2010): Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. In: *BMC bioinformatics* 11 (1), S. 567. DOI: 10.1186/1471-2105-11-567.

Batini, Carlo; Scannapieca, Monica (Hg.) (2006): Data Quality: Concepts, Methodologies and Techniques. Berlin, Heidelberg: Springer Berlin Heidelberg.

Batini, Carlo; Scannapieco, Monica (Hg.) (2016): Data and Information Quality. Cham: Springer International Publishing (Data-Centric Systems and Applications).

Bauernhansl, Thomas (2023): Die Entwicklungsstufen der Digitalen Transformation. In:, S. 3–12.

Bernhard, Jochen; Dragan, Miroslaw (2007): Bewertung der Informationsgüte in der Informationsgewinnung für die modellgestützte Analyse großer Netze der Logistik. Unter Mitarbeit von Technische Universität Dortmund.

Berry, Michael J. A.; Linoff, Gordon S. (2009): Data mining techniques: John Wiley & Sons.

Biswas, Ayan; Dutta, Soumya; Turton, Terece L.; Ahrens, James (2022): Sampling for Scientific Data Analysis and Reduction. In: Hank Childs, Janine C. Bennett und Christoph Garth (Hg.): In Situ Visualization for Computational Science. Cham: Springer International Publishing (Mathematics and Visualization), S. 11–36.

Bodendorf, Freimut (2006): Daten- und Wissensmanagement. 2., aktualisierte und erw. Aufl. Berlin, Heidelberg, New York: Springer (Springer-Lehrbuch).

Borrohou, Sanae; Fissoune, Rachida; Badir, Hassan (2025): Critical Role of Data Transformation in Preprocessing: Methods, Algorithms, and Challenges. In: Carlos Ordonez, Giancarlo Sperlì, Elio Masciari und Ladjel Bellatreche (Hg.): Model and Data Engineering, Bd. 15590. Cham: Springer Nature Switzerland (Lecture Notes in Computer Science), S. 108–122.

Böttiger, Werner; Chamoni, Peter; Gluchowski, Peter; Müller, Jochen (2001): Ein Kriterienkatalog zur Beurteilung und Einordnung von Data Warehouse-Lösungen. In: Wolfgang Behme und Harry Mucksch (Hg.): Data Warehouse-gestützte Anwendungen: Theorie und Praxiserfahrungen in verschiedenen Branchen. Wiesbaden: Gabler Verlag, S. 33–58.

Brewer, Eric A. (2000): Towards robust distributed systems (abstract). In: Gil Neiger (Hg.): Proceedings of the nineteenth annual ACM symposium on Principles of distributed computing. PODC00: ACM Symposium on Principles of Distributed Computing. Portland Oregon USA, 16 07 2000 19 07 2000. New York, NY, USA: ACM, S. 7.

Cai, Li; Zhu, Yangyong (2015): The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. In: *CODATA* 14 (0), S. 2. DOI: 10.5334/dsj-2015-002.

Campos, Cintia Isabel de; Pitombo, Cira Souza; Delhomme, Patricia; Quintanilha, José Alberto (2020): Comparative analysis of data reduction techniques for questionnaire validation using self-reported driver behaviors. In: *Journal of safety research* 73, S. 133–142. DOI: 10.1016/j.jsr.2020.02.004.

Chakraborty, Subrata; Yeh, Chung-Hsing (2007): A Simulation Based Comparative Study of Normalization Procedures in Multiattribute Decision Making.

Challa, Jagat Sesh; Goyal, Navneet; Sharma, Amogh; Sreekumar, Nikhil; Balasubramaniam, Sundar; Goyal, Poonam (2024): A Survey and Experimental Review on Data Distribution Strategies for Parallel Spatial Clustering Algorithms. In: *J. Comput. Sci. Technol.* 39 (3), S. 610–636. DOI: 10.1007/s11390-024-2700-0.

Chen; Chiang; Storey (2012): Business Intelligence and Analytics: From Big Data to Big Impact. In: *MIS Quarterly* 36 (4), S. 1165. DOI: 10.2307/41703503.

Chu, Mui Kim; Yong, Kevin Ow (2021): Big Data Analytics for Business Intelligence in Accounting and Audit. In: *JSS* 09 (09), S. 42–52. DOI: 10.4236/jss.2021.99004.

Cleve, Jürgen; Lämmel, Uwe (2024): Data Mining. Datenanalyse für Künstliche Intelligenz. 4. Auflage. Berlin, Boston: De Gruyter (De Gruyter Studium). Online verfügbar unter https://www.degruyter.com/isbn/9783111387260.

Collins, Christopher J.; Smith, Ken G. (2006): Knowledge Exchange and Combination: The Role of Human Resource Practices in the Performance of High-Technology Firms. In: *AMJ* 49 (3), S. 544–560. DOI: 10.5465/AMJ.2006.21794671.

Dillies, M.-A.; Rau, A.; Aubert, J.; Hennequet-Antier, C.; Jeanmougin, M.; Servant, N. et al. (2013): A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. In: *Briefings in Bioinformatics* 14 (6), S. 671–683. DOI: 10.1093/bib/bbs046.

Doan, AnHai (2012): Principles of data integration. Unter Mitarbeit von Alon Halevy und Zachary G. Ives. Waltham, MA: Morgan Kaufmann. Online verfügbar unter https://learning.oreilly.com/library/view/-/9780124160446/?ar.

Dong, Xin Luna; Srivastava, Divesh (2015): Big Data Integration. Cham: Springer International Publishing.

Düsing, Roland (2006): Knowledge Discovery in Databases. In: Peter Chamoni und Peter Gluchowski (Hg.): Analytische Informationssysteme. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 241–262.

English, Larry P. (1999): Improving data warehouse and business information quality. Methods for reducing costs and increasing profits. New York: J. Wiley & Sons. Online verfügbar unter https://learning.oreilly.com/library/view/-/9780471253839/?ar.

Fahrmeir, Ludwig; Heumann, Christian; Künstler, Rita; Pigeot, Iris; Tutz, Gerhard (2016): Statistik. Berlin, Heidelberg: Springer Berlin Heidelberg.

Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996): From Data Mining to Knowledge Discovery in Databases. In: *AI Magazine* 17 (3), S. 37–54. DOI: 10.1609/aimag.v17i3.1230.

Fernández, Alberto; López, Victoria; Galar, Mikel; Del Jesus, María José; Herrera, Francisco (2013): Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. In: *Knowledge-Based Systems* 42, S. 97–110. DOI: 10.1016/j.knosys.2013.01.018.

Floridi, Luciano (2005): The Ontological Interpretation of Informational Privacy. In: *Ethics and Information Technology* 7 (4), S. 185–200. DOI: 10.1007/s10676-006-0001-7.

Frederik Möller; Markus Spiekermann; Anja Burmann; Heinrich Pettenpohl (2017): Bedeutung von Daten im Zeitalter der Digitalisierung. Unter Mitarbeit von Michael ten Hompel, Michael Henke und Uwe Clausen.

Frické, Martin (2019): The Knowledge Pyramid: the DIKW Hierarchy. In: *KO* 46 (1), S. 33–46. DOI: 10.5771/0943-7444-2019-1-33.

Fujita, André; Sato, João Ricardo; Rodrigues, Leonardo de Oliveira; Ferreira, Carlos Eduardo; Sogayar, Mari Cleide (2006): Evaluating different methods of microarray data normalization. In: *BMC bioinformatics* 7, S. 469. DOI: 10.1186/1471-2105-7-469.

Gal, Michal; Rubinfeld, Daniel L. (2018): Data Standardization. In: *SSRN Journal*. DOI: 10.2139/ssrn.3326377.

García, Salvador; Derrac, Joaquín; Cano, José Ramón; Herrera, Francisco (2012): Prototype selection for nearest neighbor classification: taxonomy and empirical study. In: *IEEE transactions on pattern analysis and machine intelligence* 34 (3), S. 417–435. DOI: 10.1109/TPAMI.2011.142.

García, Salvador; Luengo, Julián; Herrera, Francisco (2015): Data Preprocessing in Data Mining. Cham: Springer International Publishing (72).

García, Salvador; Ramírez-Gallego, Sergio; Luengo, Julián; Benítez, José Manuel; Herrera, Francisco (2016): Big data preprocessing: methods and prospects. In: *Big Data Anal* 1 (1). DOI: 10.1186/s41044-016-0014-0.

Gheware, Shital; Kejkar, Anand; Tondare, Santosh (2014): Data Mining :Task, Tools, Techniques and Applications. In: *IJARCCE* 3. DOI: 10.17148/IJARCCE.2014.31003.

Ghoul, Bilel; Atitallah, Bilel Ben; Sahnoun, Salwa; Fakhfakh, Ahmed; Kanoun, Olfa (2024): Comparative Study of Data Reduction Methods in Electrical Impedance Tomography For Hand Sign Recognition. In: 2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP). 2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP). Sousse, Tunisia, 11.07.2024 - 13.07.2024: IEEE, S. 654–658.

Haerder, Theo; Reuter, Andreas (1983): Principles of transaction-oriented database recovery. In: *ACM Comput. Surv.* 15 (4), S. 287–317. DOI: 10.1145/289.291.

Han, Jiawei (2012): Data mining. Concepts and techniques. Unter Mitarbeit von Micheline Kamber und Jian Pei. 3rd ed. Waltham, MA: Morgan Kaufmann/Elsevier (Morgan Kaufmann series in data management systems). Online verfügbar unter https://learning.oreilly.com/library/view/-/9780123814791/?ar.

Hand, David J. (2001): Data Mining. In: Abdel H. El-Shaarawi und Walter W. Piegorsch (Hg.): Encyclopedia of Environmetrics: Wiley.

Hardman, Timothy James; Tapamo, Jules-Raymond (2024): A Comparative Study of Dimensionality Reduction Techniques for Satellite Image Analysis. In: Xin-She Yang, R. Simon Sherratt, Nilanjan Dey und Amit Joshi (Hg.): Proceedings of Eighth International Congress on Information and Communication Technology, Bd. 696. Singapore: Springer Nature Singapore (Lecture Notes in Networks and Systems), S. 517–528.

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009): The Elements of Statistical Learning. New York, NY: Springer New York.

Hippner, Hajo; Grieser, Lukas; Wilde, Klaus D. (2011): Data Mining – Grundlagen und Einsatzpotenziale in analytischen CRM-Prozessen. In: Hajo Hippner, Beate Hubrich und Klaus D. Wilde (Hg.): Grundlagen des CRM. Wiesbaden: Gabler, S. 783–810.

Hippner, Hajo; Merzenich, Melanie; Wilde, Klaus D. (2004): Data Mining — Grundlagen und Einsatzpotenziale im CRM. In: Klaus D. Wilde und Hajo Hippner (Hg.): IT-Systeme im CRM: Aufbau und Potenziale. Wiesbaden: Gabler Verlag, S. 241–268.

Hofmann, Thomas; Schölkopf, Bernhard; Smola, Alexander J. (2008): Kernel methods in machine learning. In: *Ann. Statist.* 36 (3). DOI: 10.1214/009053607000000677.

Hompel, Michael ten; Mallée, Torsten; Heistermann, Frauke (2017): Digitalisierung in der Logistik. Antworten auf Fragen aus der Unternehmenspraxis. In: *Bundesvereinigung Logistik BVL*, S. 15.

Hosseinzadeh, Mehdi; Azhir, Elham; Ahmed, Omed Hassan; Ghafour, Marwan Yassin; Ahmed, Sarkar Hasan; Rahmani, Amir Masoud; Vo, Bay (2023): Data cleansing mechanisms and approaches for big data analytics: a systematic study. In: *J Ambient Intell Human Comput* 14 (1), S. 99–111. DOI: 10.1007/s12652-021-03590-2.

Jain, Nikita (2013): Data mining techniques: A survey paper. In: *International Journal of Research in Engineering and Technology* 02, S. 116–119. DOI: 10.15623/ijret.2013.0211019.

Janssen, Jürgen; Laatz, Wilfried (1994): Statistische Datenanalyse mit SPSS für Windows. Berlin, Heidelberg: Springer Berlin Heidelberg.

Jolliffe, Ian (2011): Principal Component Analysis. In:, S. 1094–1096.

Kahn, Michael; Callahan, Tiffany; Barnard, Juliana; Bauck, Alan; Brown, Jeffrey; Davidson, Bruce et al. (2016): A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. In: *eGEMs (Generating Evidence & Methods to improve patient outcomes)* 4. DOI: 10.13063/2327-9214.1244.

Katharina Schüller; Paulina Busch ·; Carina Hindinger (2019): Future Skills: Ein Framework für Data Literacy.

Kemper, Alfons; Eickler, André (2015): Datenbanksysteme: De Gruyter Oldenbourg (Eine Einführung).

Kleuker, Stephan (2016): Grundkurs Datenbankentwicklung: Springer.

Knobloch, Bernd (2001): Der Data-Mining-Ansatz zur Analyse betriebswirtschaftlicher Daten. In: *Rundbrief der GI-Fachgruppe 5.10* 8 (1).

Krahl, Daniela; Windheuser, Ulrich; Zick, Friedrich-Karl (1998): Data Mining: Einsatz in der Praxis: Addison-Wesley.

Krishna, Chenchu Murali; Ruikar, Kirti; Jha, Kumar Neeraj (2023): Determinants of Data Quality Dimensions for Assessing Highway Infrastructure Data Using Semiotic Framework. In: *Buildings* 13 (4), S. 944. DOI: 10.3390/buildings13040944.

Küppers, Bertram (1999): Data mining in der Praxis: ein Ansatz zur Nutzung der Potentiale von Data mining im betrieblichen Umfeld: Lang.

Laurini (2017): Geographic knowledge infrastructure. Applications to territorial intelligence and smart cities. Amsterdam: Elsevier.

Lee, Yang W.; Pipino, Leo L.; Wang, Richard Y.; Funk, James D. (2006): Journey to Data Quality: The MIT Press.

Leskovec, Jure; Rajaraman, Anand; Ullman, Jeffrey David (2020): Mining of Massive Datasets. 3. Aufl. Cambridge: Cambridge University Press. Online verfügbar unter https://www.cambridge.org/core/product/9CFE54DDEC0DA5F9558597816E5848C3.

Lewandowski, Dirk (1999): Informationsarmut: sowohl ein quantitatives als auch ein qualitatives Problem. In: *BIBLIOTHEK Forschung und Praxis* 23 (1). DOI: 10.1515/bfup.1999.23.1.5b.

Li, Jing; Chen, Hewan; Shahizan, Mohd Othman; Yusuf, Lizawati Mi (2024): Enhancing IoT security: A comparative study of feature reduction techniques for intrusion detection system. In: *Intelligent Systems with Applications* 23, S. 200407. DOI: 10.1016/j.iswa.2024.200407.

Lieber, Daniel; Erohin, Olga; Deuse, Jochen (2013): Wissensentdeckung im industriellen Kontext. In: *Zeitschrift für wirtschaftlichen Fabrikbetrieb* 108 (6), S. 388–393. DOI: 10.3139/104.110948.

Lima, Felipe Tomazelli; Souza, Vinicius M.A. (2023): A Large Comparison of Normalization Methods on Time Series. In: *Big Data Research* 34, S. 100407. DOI: 10.1016/j.bdr.2023.100407.

Linoff, Gordon; Berry, Michael (2011): Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.

Little, Roderick; Rubin, Donald (2019): Statistical Analysis with Missing Data, Third Edition: Wiley.

Liu, Xueyan; Li, Nan; Liu, Sheng; Wang, Jun; Zhang, Ning; Zheng, Xubin et al. (2019): Normalization Methods for the Analysis of Unbalanced Transcriptome Data: A Review. In: *Frontiers in bioengineering and biotechnology* 7, S. 358. DOI: 10.3389/fbioe.2019.00358.

Maimon, Oded; Rokach, Lior (Hg.) (2010): Data Mining and Knowledge Discovery Handbook. Boston, MA: Springer US.

Manikandan, S. (2010): Data transformation. In: *Journal of pharmacology & pharmacotherapeutics* 1 (2), S. 126–127. DOI: 10.4103/0976-500X.72373.

Mâsse, Louise C.; Fuemmeler, Bernard F.; Anderson, Cheryl B.; Matthews, Charles E.; Trost, Stewart G.; Catellier, Diane J.; Treuth, Margarita (2005): Accelerometer data reduction: a comparison of four reduction algorithms on select outcome variables. In: *Medicine and science in sports and exercise* 37 (11 Suppl), S544-54. DOI: 10.1249/01.mss.0000185674.09066.8a.

Meier, Andreas; Kaufmann, Michael (2016): SQL- & NoSQL-Datenbanken. Berlin, Heidelberg: Springer Berlin Heidelberg.

Mertens, Peter; Bodendorf, Freimut; König, Wolfgang; Schumann, Matthias; Hess, Thomas; Buxmann, Peter (2017): Grundzüge der Wirtschaftsinformatik. Berlin, Heidelberg: Springer Berlin Heidelberg.

Michael Goebel; Le Gruenwald (1999): A survey of data mining and knowledge discovery software tools. In: *SIGKDD Explor* 1, S. 20–33. Online verfügbar unter https://api.semanticscholar.org/CorpusID:2007403.

Michnik, Jerzy; Lo, Mei-Chen (2009): The assessment of the information quality with the aid of multiple criteria analysis. In: *European Journal of Operational Research* 195 (3), S. 850–856. DOI: 10.1016/j.ejor.2007.11.017.

Müller, Heiko; Freytag, Johann-Christoph (2003): Problems, methods, and challenges in comprehensive data cleansing.

N. Sharma; K. Saroha (2015): Study of dimension reduction methodologies in data mining. In: International Conference on Computing, Communication & Automation. International Conference on Computing, Communication & Automation, S. 133–137.

Naisbitt, John (1982): Megatrends. Ten new directions transforming our lives. 1. printing. New York, NY: Warner Books (Warner books).

Nasser, Adel A.; Alkhulaidi, A. A.; Ali, Mansoor N.; Hankal, M.; Al-olofe, M. (2019): A Study on the impact of multiple methods of the data normalization on the result of SAW, WED and TOPSIS ordering in Healthcare Multi-attributtes Decision Making Systems based on EW, ENTROPY, CRITIC and SVP weighting approaches. In: *Indian Journal of Science and Technology* 12 (4), S. 1–21. DOI: 10.17485/ijst/2019/v12i4/140756.

North, Klaus (2016): Wissensorientierte Unternehmensführung. Wiesbaden: Springer Fachmedien Wiesbaden.

North, Klaus (2021): Die Wissenstreppe. In: Klaus North (Hg.): Wissensorientierte Unternehmensführung: Wissensmanagement im digitalen Wandel. Wiesbaden: Springer Fachmedien Wiesbaden, S. 33–69.

Osborne, Jason (2010): Improving your data transformations: Applying Box-Cox transformations as a best practice. In: *Pract Assess Res Eval* 15, S. 1–9.

Peshawa J Muhammad Ali; Rezhna Hassan Faraj (2014): Data Normalization and Standardization: A Technical Report.

Petersohn, Helge (2005): Data Mining: Oldenbourg Wissenschaftsverlag GmbH.

Probst, Gilbert; Raub, Stefan; Romhardt, Kai (2006): Wissen managen. Wie Unternehmen ihre wertvollste Ressource optimal nutzen. 5., überarb. Aufl. Wiesbaden: Gabler.

Ramírez-Gallego, Sergio; García, Salvador; Mouriño-Talín, Héctor; Martínez-Rego, David; Bolón-Canedo, Verónica; Alonso-Betanzos, Amparo et al. (2016): Data discretization: taxonomy and big data challenge. In: *WIREs Data Min & Knowl* 6 (1), S. 5–21. DOI: 10.1002/widm.1173.

Redman, Thomas C. (1998): The impact of poor data quality on the typical enterprise. In: *Commun. ACM* 41 (2), S. 79–82. DOI: 10.1145/269012.269025.

Rogova, G. L.; Bosse, E. (2010): Information quality in information fusion. In: 2010 13th International Conference on Information Fusion. 2010 13th International Conference on Information Fusion (FUSION 2010). Edinburgh, 26.07.2010 - 29.07.2010: IEEE, S. 1–8.

Röhner, Jessica; Schütz, Astrid (2016): Klassische Kommunikationsmodelle. In: Jessica Röhner und Astrid Schütz (Hg.): Psychologie der Kommunikation. Wiesbaden: Springer Fachmedien Wiesbaden, S. 19–38.

Rohweder, Jan P.; Kasten, Gerhard; Malzahn, Dirk; Piro, Andrea; Schmid, Joachim (2008): Informationsqualität — Definitionen, Dimensionen und Begriffe. In: Knut Hildebrand, Marcus Gebauer, Holger Hinrichs und Michael Mielke (Hg.): Daten- und Informationsqualität: Auf dem Weg zur Information Excellence. Wiesbaden: Vieweg+Teubner, S. 25–45.

Rudy, Jason; Valafar, Faramarz (2011): Empirical comparison of cross-platform normalization methods for gene expression data. In: *BMC bioinformatics* 12, S. 467. DOI: 10.1186/1471-2105-12-467.

Runkler, Thomas A. (2010): Data Mining. Wiesbaden: Vieweg+Teubner.

Ryu, Kyung-Seok; Park, Joo-Seok; Park, Jae-Hong (2006): A data quality management maturity model. In: *ETRI journal* 28 (2), S. 191–204.

S. K. Joshi; S. Machchhar (2014): An evolution and evaluation of dimensionality reduction techniques — A comparative study. In: 2014 IEEE International Conference on Computational Intelligence and Computing Research. 2014 IEEE International Conference on Computational Intelligence and Computing Research, S. 1–5.

Schelter, Sebastian; Lange, Dustin; Schmidt, Philipp; Celikel, Meltem; Biessmann, Felix; Grafberger, Andreas (2018): Automating large-scale data quality verification. In: *Proc. VLDB Endow.* 11 (12), S. 1781–1794. DOI: 10.14778/3229863.3229867.

Schicker, Edwin (2017): Datenbanken und SQL. Wiesbaden: Springer Fachmedien Wiesbaden.

Schmidt, Marcin T.; Handschuh, Luiza; Zyprych, Joanna; Szabelska, Alicja; Olejnik-Schmidt, Agnieszka K.; Siatkowski, Idzi; Figlerowicz, Marek (2011): Impact of DNA microarray data transformation on gene expression analysis - comparison of two normalization methods. In: *Acta Biochim Pol* 58 (4). DOI: 10.18388/abp.2011_2227.

Sellami, Akrem; Farah, Mohamed (2018): Comparative study of dimensionality reduction methods for remote sensing images interpretation. In: 2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). 2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). Sousse, 21.03.2018 - 24.03.2018: IEEE, S. 1–6.

Sharma, Vinod (2022): A Study on Data Scaling Methods for Machine Learning. In: *INJGAADMIN, ftjijgasr* 1 (1). DOI: 10.55938/ijgasr.v1i1.4.

Silberschatz, Abraham; Korth, Henry F.; Sudarshan, S. (2020): Database system concepts. Seventh edition. New York, NY: McGraw-Hill Education.

Song, M.; Yang, H.; Siadat, S. H.; Pechenizkiy, M. (2013): A comparative study of dimensionality reduction techniques to enhance trace clustering performances. In: *Expert Systems with Applications* 40 (9), S. 3722–3737. DOI: 10.1016/j.eswa.2012.12.078.

Steiner, René (2021): Grundkurs Relationale Datenbanken. Wiesbaden: Springer Fachmedien Wiesbaden.

Tanenbaum, Andrew S.; Chandavarkar, B. R.; Austin, Todd (2013): Structured computer organization. Sixth edition. Harlow, Boston: Pearson Education Limited. Online verfügbar unter https://elibrary.pearson.de/book/99.150005/9780273775331.

Thearling, Kurt (2010): Data Mining for CRM. In: Oded Maimon und Lior Rokach (Hg.): Data Mining and Knowledge Discovery Handbook. Boston, MA: Springer US, S. 1181–1188.

Tiwari, S.: Professional Nosql: Wiley India Pvt. Limited. Online verfügbar unter https://books.google.de/books?id=ECl0CgAAQBAJ.

Uhm, Daiho; Jun, Sung-Hae; Lee, Seung-Joo (2012): A Classification Method Using Data Reduction. In: *International Journal of Fuzzy Logic and Intelligent Systems* 12 (1), S. 1–5. DOI: 10.5391/IJFIS.2012.12.1.1.

van der Maaten, Laurens; Postma, Eric; Herik, H. (2007): Dimensionality Reduction: A Comparative Review. In: *Journal of Machine Learning Research - JMLR* 10.

Vieweg, Iris; Werner, Christian; Wagner, Klaus-P.; Hüttl, Thomas; Backin, Dieter (2012): Datenbanken. In: Klaus-P. Wagner, Thomas Hüttl, Dieter Backin, Iris Vieweg und Christian Werner (Hg.): Einführung Wirtschaftsinformatik. Wiesbaden: Gabler Verlag, S. 103–129.

Villares, Mario; Saunders, Carla M.; Fey, Natalie (2024): Comparison of dimensionality reduction techniques for the visualisation of chemical space in organometallic catalysis. In: *Artificial Intelligence Chemistry* 2 (1), S. 100055. DOI: 10.1016/j.aichem.2024.100055.

Vu, Thao; Riekeberg, Eli; Qiu, Yumou; Powers, Robert (2018): Comparing normalization methods and the impact of noise. In: *Metabolomics : Official journal of the Metabolomic Society* 14 (8), S. 108. DOI: 10.1007/s11306-018-1400-6.

Wagner, Klaus-P.; Hüttl, Thomas; Backin, Dieter; Vieweg, Iris; Werner, Christian (Hg.) (2012): Einführung Wirtschaftsinformatik. Wiesbaden: Gabler Verlag.

Wang, Richard Y.; Strong, Diane M. (1996): Beyond Accuracy: What Data Quality Means to Data Consumers. In: *Journal of Management Information Systems* 12 (4), S. 5–33. Online verfügbar unter http://www.jstor.org/stable/40398176.

Wang, Zihao; Zhang, Guiyong; Xing, Xiuqing; Xu, Xiangguo; Sun, Tiezhi (2024): Comparison of dimensionality reduction techniques for multi-variable spatiotemporal flow fields. In: *Ocean Engineering* 291, S. 116421. DOI: 10.1016/j.oceaneng.2023.116421.

Witten, Ian H.; Frank, Eibe; Hall, Mark A.; Pal, Christopher J. (2017): Data mining. Practical machine learning tools and techniques. Fourth edition. Amsterdam, Boston, Heidelberg, London, New York, Oxford, Paris, San Diego, San Francisco, Singapore, Sydney, Tokyo: Elsevier Morgan Kaufmann (Morgan Kaufmann series in data management systems). Online verfügbar unter https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=4708912.

Wöstmann, R.; Strauß, P.; Deuse, J. (2017): Predictive Maintenance in der Produktion. In: *Anwendungsfälle und Einführungsvoraussetzungen zur Erschließung ungenutzter Potentiale. Werkstattstechnik online* 107 (7/8), S. 524–529.

Xu, Rui; Wunsch, Donald (2005): Survey of Clustering Algorithms. In: *Neural Networks, IEEE Transactions on* 16, S. 645–678. DOI: 10.1109/TNN.2005.845141.

Yang, Xin-She (2024): Proceedings of Eighth International Congress on Information and Communication Technology. ICICT 2023, London, Volume 4. Unter Mitarbeit von R. Simon Sherratt, Nilanjan Dey und Amit Joshi. 1st ed. Singapore: Springer (Lecture Notes in Networks and Systems Series, v.696). Online verfügbar unter https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=30745214.

Yesilyurt, Ozan; Theologou, Alexander; Schlereth, Andreas (2023): Cloudbasiertes Intelligentes C-Teile-Management. In:, S. 343–354.

You, Shingchern D.; Hung, Ming-Jen (2021): Comparative Study of Dimensionality Reduction Techniques for Spectral–Temporal Data. In: *Information* 12 (1), S. 1. DOI: 10.3390/info12010001.

Zebari, Rizgar; Abdulazeez, Adnan; Zeebaree, Diyar; Zebari, Dilovan; Saeed, Jwan (2020): A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. In: *JASTT* 1 (1), S. 56–70. DOI: 10.38094/jastt1224.

Zins, Chaim (2007): Conceptual approaches for defining data, information, and knowledge. In: *J. Am. Soc. Inf. Sci.* 58 (4), S. 479–493. DOI: 10.1002/asi.20508.

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

CRISP-DM	Cross Industry Standard Process for Data Mining
DBMS	Datenbankmanagementsystem
FA	Faktoranalyse
ICA	Independent Component Analysis
k-NN	k-Nearest Neighbour
KE	Korrelationserhaltung
KPCA	Kernel-PCA
LDA	Lineare Diskriminanzanalyse
MRMR	Minimum Redundancy Maximum Relevance
MI	Merkmalsinterdependenz
NV	Normalitätsverbesserung
PCA	Hauptkomponentenanalyse
RUL	Restnutzungsdauer
SG	Streuungsnutzungsgrad
SHI	Index zur Skalierungsheterogenität
SV	Symmetrieverbesserung
SVM	Support-Vector-Machine
VDBS	Verteiltes Datenbanksystem

Anhang A: Liste statistischer Kennzahlen

Name	Formel	Definition	Quelle
Variationskoeffizient	$v = \frac{s}{\bar{x}}$	v: empirischer Variationskoeff. s: empirische Varianz $\bar{x}$: Mittelwert	Böselt, M. (2018) *Statistik*. Berlin: Oldenbourg Wissenschaftsverlag
Mittelwert Standardabw.	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$	σ: empirische Standardabweichun $\bar{x}$: Mittelwert	Böselt, M. (2018) *Statistik*. Berlin: Oldenbourg Wissenschaftsverlag
Shapiro Wilk Schätzung	$W = \frac{b^2}{(n-1)s^2}$	W: Shapiro Teststatistik b^2: Schätzer 1 s^2: Schätzer 2	Shapiro, S.S. and Wilk, M.B. (1965) 'An analysis of variance test for normality (complete samples)', *Biometrika*, 52(3/4), p. 591. doi:10.2307/2333709.
Fischer-Pearson Schiefe bzw. empirischer Korrelationskoeffizient	$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$	r_{XY}: empirischer Korrelationskoeffiz zu einer Zahl x und y	Böselt, M. (2018) *Statistik*. Berlin: Oldenbourg Wissenschaftsverlag
Konditionszahl	$\kappa_{abs} = \lim_{\tilde{x}\to x} \sup \frac{\|f(\tilde{x}) - f(x)\|}{\|\tilde{x} - x\|}$	κ_{abs}: absolute Konditionszahl x: Eingabewert $\tilde{x}$: gestörter Eingabewert	Deuflhard, P. and Hohmann, A. 2019. *Numerische Mathematik 1: Eine algorithmisch orientierte Einführung*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110614329
Mutual information (Transinformation)	$I(X;Y) = \sum_x \sum_y p(x,y) log_2 (\frac{p(x,y)}{p(x)q(y)})$	$p(x), q(y)$: Auftrittswahrscheinlic	Rinne, Horst (2008) Taschenbuch der *Statistik*. 4. Auflage. Harri Deutsch, ISBN 978-3-8171-1827-4, S. 64
Spearman Rangkoeffizient	$\widehat{\rho_r} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$	$\widehat{\rho_r}$: empirischer Rang − Korrelationskoeffizient d_i: Rangzahl differenz	Böselt, M. (2018) *Statistik*. Berlin: Oldenbourg Wissenschaftsverlag

Silhouetten-Score	$S(o) = \begin{cases} 0 \text{ , für wenn o einziges Elemen} \\ \frac{dist(B,o) - dist(A,o)}{\max\{dist(A,o), dist(B,o)\}}, \end{cases}$	$dist: Distance$	Rousseeuw, P.J. (1987) 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, 20, pp. 53–65. doi:10.1016/0377-0427(87)90125-7.

Anhang B: Metriken

Metriken	Formel	Referenz
Informationserhaltung	$IE = \frac{I_{nach}}{I_{vor}}$	(2)
Statistische Ähnlichkeit	$VE_k = \frac{\sum_{i=1}^{k} \mathrm{Var}_i^{\mathrm{nach}}}{\sum_{j=1}^{k} \mathrm{Var}_j^{\mathrm{vor}}}$	(1)
	$KE = 1 - \lvert\rho_{vor} - \rho_{nach}\rvert$	(3)
Skalierungsanpassung	$SG = \mu\left(\frac{\sigma_i}{\max(X_i) - \min(X_i)}\right)$	(7)
	$SHI = \frac{\sigma_{Spannweite}}{\mu_{Spannweite}}$	(6)
	Cluster-Trennbarkeit: Vergleich Silhouetten-Score	
Verteilungsoptimierung	$NV = \frac{p_{nach} - p_{vor}}{1 - p_{vor}}$	(8)
	$SV = \frac{\lvert skew_{vor}\rvert - \lvert skew_{nach}\rvert}{\lvert skew_{vor}\rvert}$	(9)
Numerische Stabilität	$\kappa(X) = \frac{\lambda_{\max}(X)}{\lambda_{\min}(X)}$	(10)
	$MI = \mu(\lvert\rho_{ij}\rvert)$, für $i \neq j$	(4)
Ressourceneffizienz	Einfache Zeitmessung	
Modellleistung	Einfache Messung des Ressourcen-bedarf	
Sonstige	$Spannweite = \max(X_i) - \min(X_i)$	(5)