technische universität dortmund

IT in Produktion und Logistik

# Systematische Untersuchung von Datentransformationen im Kontext der Datenvorverarbeitung für Datenbestände aus Produktion und Logistik

Bachelorarbeit zur Erlangung des Grades Bachelor of Science (B. Sc.)


Vorgelegt von: Michael Bröker

Matrikelnummer: 186982

Studiengang: Logistik



Ausgabedatum: 19.09.2023

Abgabedatum: 12.12.2023



Erstprüfer:
Dr.-Ing. Dipl.-Inf. Anne Antonia Scheidler

Zweitprüfer:
M. Sc. Florian Hochkamp

Technische Universität Dortmund

Fakultät Maschinenbau

Fachgebiet IT in Produktion und Logistik

# Inhaltsverzeichnis

Abbildungsverzeichnis ..... II
Tabellenverzeichnis ..... III
Abkürzungsverzeichnis ..... IV
1 Einleitung ..... 1
2 Wissensentdeckung in Produktion und Logistik ..... 3
2.1 Datentransformation im Kontext der Datenvorverarbeitung ..... 3
2.1.1 Daten und Wissen ..... 3
2.1.2 Vorgehensmodelle zur Wissensentdeckung ..... 5
2.1.3 Datenvorverarbeitung ..... 7
2.1.4 Datentransformation ..... 9
2.2 Produktion und Logistik ..... 11
2.2.1 Produktion als Wertschöpfungsprozess ..... 11
2.2.2 Grundlagen Logistik ..... 13
3 Untersuchung von Datentransformation ..... 14
3.1 Methodik der systematischen Literaturrecherche ..... 14
3.2 Datentransformation zur Knowledge Discovery in Databases ..... 18
3.3 Datentransformationsverfahren ..... 22
3.3.1 Standardisierung und Normalisierung ..... 25
3.3.2 Datenglättung ..... 28
3.3.3 Diskretisierung ..... 31
3.4 Diskussion und Fazit ..... 33
4 Zusammenfassung und Ausblick ..... 36
Literaturverzeichnis ..... 37
Anhang ..... 40
Anhang A: Genutzte Quellen zum Verständnis der Datentransformation ..... 40
Eidesstattliche Versicherung ..... 42

# Abbildungsverzeichnis

Abbildung 1 Hierarchie Zeichen, Daten, Informationen und Wissen ..... 3
Abbildung 2 Kategorien von Datentypen ..... 4
Abbildung 3 KDD-Stufenmodell nach Fayyad ..... 5
Abbildung 4 Crisp-DM Modell ..... 6
Abbildung 5 Aufbau eine Arbeitssystems ..... 11
Abbildung 6 Aufbau strukturierte Literaturrecherche ..... 14
Abbildung 7 Aufbau Datenvorbereitung zur Knowledge Discovery ..... 20
Abbildung 8 Aufbau Datenvorverarbeitung zur Knowledge Discovery ..... 21
Abbildung 9 Übersicht der Methoden zur Datentransformation in Bezug zur Kategorie Transformation ..... 24
Abbildung 10 Normalisierung eines Datensatzes ..... 26
Abbildung 11 Beispielhafte Clusterbildung ..... 29
Abbildung 13 Kategorien zu Diskretisierung ..... 33

# Tabellenverzeichnis

Tabelle 1 Taxonomie nach Cooper.....15
Tabelle 2 Vergleich der bekanntesten KDD-Modelle .....18
Tabelle 3 Tabellenauszug der genannten Verfahren zur Datentransformation .....19
Tabelle 4 Tabbelenauszug Standardisierung und Normalisierung .....25
Tabelle 5 Quellenübersicht zur Datenglättung .....28
Tabelle 6 Tabellenauszug zur Diskretisierung .....31

## Abkürzungsverzeichnis

| | |
|---|---|
| KDD | Knowledge Discovery in Databases |
| CRISP-DM | Cross Industry Standard Process for Data Mining |
| SV | supervised (überwacht |
| USV | unsupervised (unüberwacht |

# 1 Einleitung

Daten sind heutzutage in sämtlichen Branchen allgegenwärtig (Bensberg 2001). Die zunehmende Digitalisierung und der verstärkte Einsatz von Informationstechnologien haben zu einem exponentiellen Wachstum von Daten geführt (Bensberg 2001). Dies trifft daher auch auf die Produktion und Logistik zu. Immer mehr Abläufe und Vorgänge können in Echtzeit erfasst und digital gespeichert werden, um Zustände und ihre Veränderungen festzuhalten. (Müller und Lenz 2013). Durch diese Entwicklung erzeugen Unternehmen mehrere Petabytes an Daten und speichern diese ab (Luengo et al. 2020). Anhand dieser Daten werden Informationen gesammelt, welche zu zweckorientiertem Wissen führen, um betriebliche Entscheidungen zu treffen (Schuh et al. 2022, S. 18). Das Datenvolumen, sowie die Verfügbarkeit von Daten und das damit verbundene Wissen, welches einem Unternehmen zur Verfügung steht, hat sich als eigener Produktionsfaktor etabliert und dient mittlerweile als wesentliche Grundlage für die Unternehmensbewertung an der Börse (Schuh et al. 2022; Brauckmann 2019). Daher stellen Daten einen wertvollen Rohstoff dar. Allerdings müssen aus den Daten zunächst Muster erkannt werden, um daraus Wissen zu identifizieren (Fayyad et al. 1996). Hierzu ist eine angemessene Datenvorverarbeitung, einschließlich der Datentransformation, für Datenbestände von entscheidender Bedeutung (Sharafi 2013).

Die Datenvorverarbeitung ist eine Phase zur Wissensentdeckung (Sharafi 2013), Zum Entdecken von Wissen in Datenbanken wurden in den vergangenen Jahren verschieden Modelle aufgestellt. Die zwei bekanntesten sind davon sind das Knowledge Discovery in Databases Stufenmodell (KDD-Stufenmodell) von Fayyad et al. (1996), sowie das Crisp DM-Modell von Wirth und Hipp (2000). Diese Modelle unterscheiden sich in ihrem Aufbau darin, dass sie die Datentransformation unterschiedlich einordnen. Im KDD-Stufenmodell nach Fayyad et al. (1996) stellt die Datentransformation eine eigene Phase dar, welche den Datensatz an die folgenden Data Mining-Verfahren anpasst. Dieses Ziel wird auch im Modell Cross Industry Standard Process for Data Mining (CRISP-DM) von Wirth und Hipp (2000) verfolgt. Allerdings sehen die Autoren die Datentransformation als einen Teil der Datenvorverarbeitung an. Durch diese unterschiedlichen Einordnungen werden der Datentransformation unterschiedliche Verfahren zugeordnet, weswegen die Datentransformation nicht klar definiert ist.

Das Ziel dieser Bachelorarbeit ist ein Verständnis des Begriffes Datentransformation im Kontext der Datenvorverarbeitung für Datenbestände in Produktion und Logistik zu erlangen. Um dieses Ziel zu erreichen, wird die Bachelorarbeit in drei Teilziele unterteilt. Das erste Teilziel umfasst die Durchführung einer systematischen Literaturrecherche zur Datentransformation, um diese zu erfassen. Als zweites Teilziel sind die Gemeinsamkeiten und Unterschiede der Datentransformation mittels eines exemplarischen Vergleichs zu untersuchen. Auf Basis der vorangegangenen Teilziele umfasst das dritte Teilziel die Definition des Begriffs Datentransformation, um so das Verständnis zu erlangen.

Um das Ziel der Arbeit zu erreichen und ein Verständnis der Datentransformation im Kontext der Datentransformation für Datenbestände zu erreichen werden zunächst die Grundlagen betrachtet. Hierzu wird im zweiten Kapitel mit der Definition von Daten und der damit verbundenen Abgrenzung von Daten zu Information und Wissen eingeführt. Dies umfasst zusätzlich die Darstellung der verschiedenen Datentypen. Darauf aufbauend werden die zwei bekanntesten Vorgehensmodelle zur Erlangung von Wissen in Datenbanken (Knowledge Discovery in Databases) vorgestellt. Dies umfasst das KDD-Stufenmodell nach Fayyad et al. (1996), sowie das Crisp DM-Modell nach Wirth und Hipp (2000). Da in dieser Arbeit die Datentransformation im Kontext der Datenvorverarbeitung näher untersucht werden soll, wird im dritten Teil des ersten Unterkapitels auf die Verfahren in der Datenvorverarbeitung eingegangen, um daraufhin bestehende Verfahren der Datentransformation vorzustellen Durch diese Betrachtung wird die Grundlage zu möglichen Suchkriterien gelegt. Nachdem das Grundlegende Wissen zur Datentransformation erlangt wurde, wird im zweiten Teil des zweiten Kapitels die Domäne

Produktion und Logistik vorgestellt. Um diese besser zu verstehen, wird zunächst die Produktion als Wertschöpfungsprozess dargestellt. Infolgedessen werden die grundlegenden Prozesse der Logistik näher betrachtet. Das dritte Kapitel leitet in den Hauptteil der Arbeit ein. Hierzu wird zunächst die Systematische Literaturrecherche nach vom Brocke et al (2009) vorgestellt. Darauf aufbauend wird das Vorgehen der systematischen Literaturrecherche dargestellt. In diesem Teil wird der Umfang und der Themenbereich der Recherche eingegrenzt, woraus sich die Suchkriterien ableiten. Im Anschluss wird dargestellt, wie die Literaturrecherche durchgeführt und analysiert wurde. Anhand der gefundenen Literatur wird ein exemplarischer Vergleich durchgeführt, durch den die gefundenen Verfahren der Datentransformation zugeordnet werden, um diese daraufhin in der Knowledge Discovery einzuordnen. Anhand der gefundenen Ergebnisse wird die Datentransformation im dritten Unterkapitelteil dargestellt und abschließend definiert. Um ein tieferes Verständnis zu erhalten, stellen die Unterkapitel aus diesem Teil die genutzten Verfahren näher vor. Zum Abschluss werden die wichtigsten Punkte der Arbeit zusammengefasst und ein Ausblick gegeben.

# 2 Wissensentdeckung in Produktion und Logistik

In den Nachfolgenden Unterkapiteln wird auf die Grundlagen dieser Arbeit eingegangen. Zunächst werden Daten näher beleuchtet. Dazu erfolgt eine Einordnung, sowie eine Definition von Daten im Kontext von Informationen und Wissen in Kapitel 2.1.1. Um Datentransformation besser zu verstehen, wird in Kapitel 2.1.2 auf die bekanntesten Vorgehensmodelle zur Wissensentdeckung eingegangen, um daraufhin in den beiden Unterkapiteln die Datenvorverarbeitung, sowie die Datentransformation näher zu betrachten. In der Literatur werden die Begriffe Data Preprocessing und Data Preparation genannt, welche sich im Deutschen unterschiedlich übersetzt werden. Im CRISP-DM Model Datenvorbereitung als Phase definiert. Allerdings übernimmt ein Großteil der Literatur, sowie der Titel dieser Arbeit die Datenvorverarbeitung als Phase, weswegen diese Definition weiter verfolgt wird.

## 2.1 Datentransformation im Kontext der Datenvorverarbeitung

### 2.1.1 Daten und Wissen

Der Begriff Daten wurde in der Literatur bereits intensiv untersucht, weswegen es bereits eine Vielzahl an Publikationen gibt, welche sich mit der Definition von Daten auseinandersetzten. Die Literatur unterscheidet zwischen Zeichen, Daten, Informationen und Wissen (North 2011; Bodendorf 2006). Wie diese hierarchisch zusammenhängen, wird in Abbildung 2 nach (Bodendorf 2006) dargestellt.

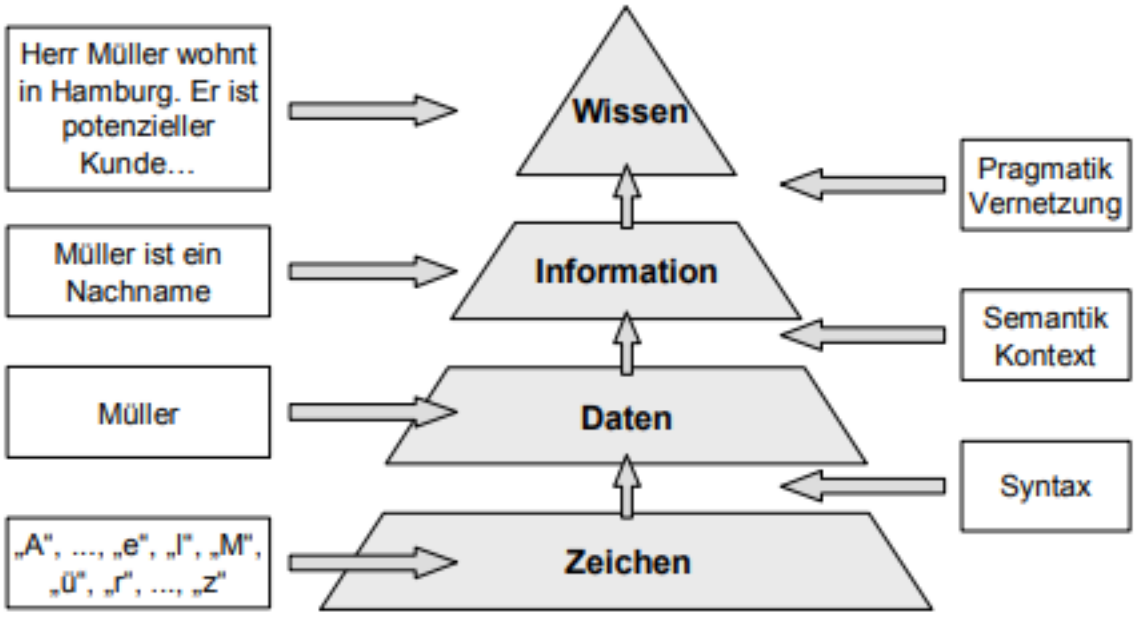


*Abbildung 1 Hierarchie Zeichen, Daten, Informationen und Wissen*

In diesem Zusammenhang sind Zeichen die unterste Ebene der Hierarchie, wobei Zeichen einzelne Buchstaben, Ziffern oder Sonderzeichen sein können. Durch eine Struktur dieser Zeichen gemäß bestimmter Syntaxregeln entstehen Daten (North 2011). Nach (Bodendorf 2006) werden Daten erst dann zu Informationen, wenn ihnen eine Bedeutung (Semantik) zugeordnet wird. Diese Zuordnung erfolgt, indem die Daten mit einem Begriff oder einer Vorstellung aus der realen oder theoretischen Welt in Verbindung gebracht werden. Fehlt beispielsweise diese Verbinndung, können Unternehmen die Daten nicht unmittelbar verwenden (Dommermuth 2020). Um aus den Informationen Wissen zu generieren, erfordert es ein Verständnis darüber, wie die verschiedenen Informationen miteinander interagieren können (Bodendorf 2006). In-

formationen müssen daher miteinander verknüpft und in einen Kontext gebracht werden (Dommermuth 2020; Bodendorf 2006). Cleve und Lämmel (2020) definieren Wissen als die Fähigkeit Informationen zu nutzen. Wissen ist also immer an individuelle Personen gebunden und hängt von deren bereits vorhandenem Wissen ab, da Informationen unterschiedlich interpretiert und genutzt werden können.

Daher ist die Betrachtung von Daten als Informationseinheit eine gültige Interpretation (Cleve und Lämmel 2020). Hier wird in unstrukturierte, semiunstrukturierte und strukturierte Daten unterschieden (Cleve und Lämmel 2020). Unstrukturierte Daten können Bilder und Texte sein (Cleve und Lämmel 2020). Diese unstrukturierte Daten können allerdings durch eine äußere Struktur, wie Beispielsweise in einer Webseite in Form gebracht werden, wodurch sie als semistrukturiert betrachtet werden können (Cleve und Lämmel 2020). Unter strukturierten Daten verstehen Cleve und Lämmel (2020) Dateiformate, welche eine feste Struktur aufweisen. In den Datensätzen besitzen die Daten eine feste Reihenfolge, Datentypen und definierte Attribute (Cleve und Lämmel 2020). Hierzu zählen beispielsweise relationale Datenbanken oder CSV-Formate, welche durch einen Editor bearbeitet werden können (Cleve und Lämmel 2020). Daten können in verschiedenen Formen auftreten. Gemäß der Syntax der Zeichenfolge innerhalb der Daten, werden Daten in verschiedene Datentypen unterschieden. Auf oberster Ebene werden Daten nach numerischen (Ziffern), alphabetischen (Buchstaben) und alphanumerischen Daten (Ziffern, Buchstaben und Sonderzeichen) unterschieden (Mertens et al. 2012).

Nach Cleve und Lämmel (2020) können Daten anhand der Kriterien Ordnung und Rechnen kategorisiert werden An Zahlen können Rechenoperatoren angewendet werden, sie können miteinander verglichen oder geordnet werden (Cleve und Lämmel 2020). Alphabetische Daten können, wie Zahlen durch ihre Form geordnet werden. Allerdings gibt es Formen, welche nicht geordnet werden können (Cleve und Lämmel 2020). Diese Kategorien werden in Abbildung 2 dargestellt.

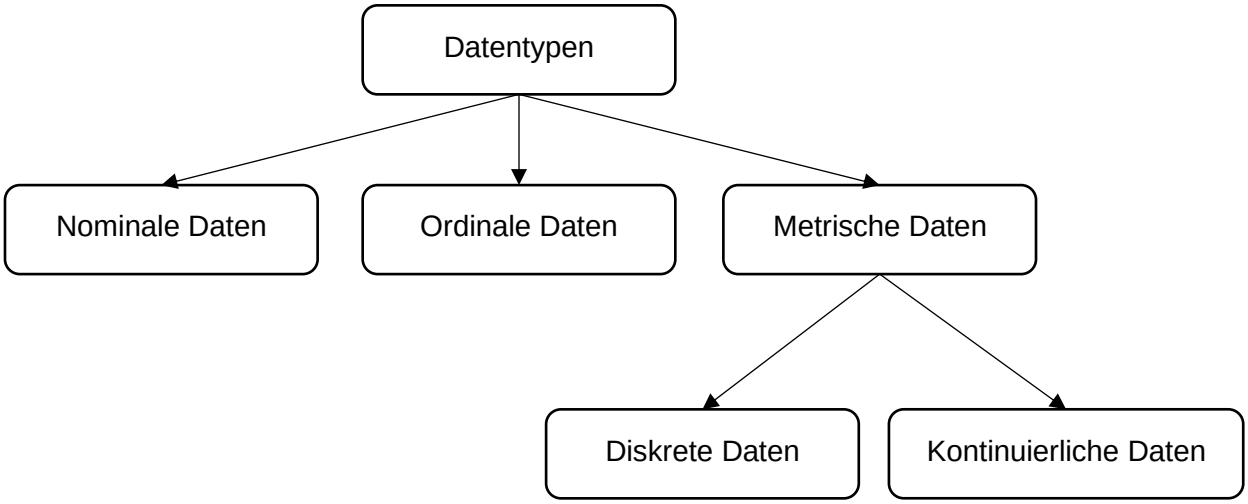


*Abbildung 2 Kategorien von Datentypen*

Mit **Nominale Daten** und können keine Berechnungen vorgenommen werden und weisen keine Reihenfolge auf (Cleve und Lämmel 2020). Daher können Sie nur dahingehend verglichen werden, ob Sie gleich bzw. nicht gleich sind (Cleve und Lämmel 2020). Hierzu gehören beispielsweise Autohersteller.

**Ordinale Daten** können im Gegensatz zu den nominalen Daten durch Operatoren wie beispielsweise größer oder kleiner geordnet werden (Cleve und Lämmel 2020). Berechnungen sind allerdings nicht möglich (Cleve und Lämmel 2020). Hierzu gehören Schulnoten (sehr gut, gut, etc) oder Körperbeschreibungen (Groß, Mittel, Klein).

Mit den **Metrischen Daten** kann sowohl gerechnet als auch Ordnungsmerkmale angewendet werden (Cleve und Lämmel 2020). Hierunter sind gehören **diskrete Daten**, welche nur Schrittweise ihre Größe verändern (Cleve und Lämmel 2020). **Kontinuierliche Daten** können hingegen jeden reellen Zahlenwert annehmen (Cleve und Lämmel 2020).

### 2.1.2 Vorgehensmodelle zur Wissensentdeckung

Zur Erlangung des Wissens aus großen Datenbeständen ist der Bedarf an computergestützten Techniken und Methoden gestiegen (Sharafi 2013). Um das angestrebte Wissen zu erlangen, ist zudem eine systematische und geplante Herangehensweise notwendig (Sharafi 2013). Die Vorgehensweisen der Wissensentdeckung in Datenbanken bzw. Knowledge Discovery in Databases (KDD) haben das Erkennen von unbekannten, nützlichen und nachvollziehbaren Mustern zum Ziel (Fayyad et al. 1996). KDD ist daher ein Prozess der Wissensidentifikation (Fayyad et al. 1996). In der Literatur wurden zu diesem Zweck viele Vorgehensmodelle veröffentlicht. Zu den bekanntesten Vorgehensmodellen gehören Cross Industry Standard Process for Data Mining (CRISP-DM) und das KDD-Stufenmodell (Kurgan und Musilek 2006). Das CRISP-DM-Modell ist aus der industriellen Praxis entstandenes Vorgehensmodell, wohingegen das KDD-Stufenmodell eine wissenschaftliche Zielgruppe anspricht (Sharafi 2013). Nach Sharafi (2013) lassen sich Vorgehensmodelle für das KDD auf 4 Phasen zusammenfassen: Die (Daten-)Vorbereitung, Datenvorverarbeitung, Methodenanwendung und die Ergebnisinterpretation. Während der Vorbereitung werden Datenquellen angeschlossen und selektiert (Sharafi 2013). Die Vorverarbeitung dient der Bereinigung, sowie der Transformation der Daten in nutzbare Formate (Sharafi 2013). Darauf aufbauend, können Methoden angewendet und die Ergebnisse mit den entsprechenden Experten des Anwendungsgebietes interpretiert werden (Sharafi 2013). Der Aufwand innerhalb dieser Phasen ist unterschiedlich groß. Nach Kurgan und Musilek (2006) beträgt der Anteil für die Datenvorverarbeitung ca. 50 bis 60 Prozent, wohingegen das eigentliche Data Mining nur 10 bis 18 Prozent des Aufwandes ausmacht.

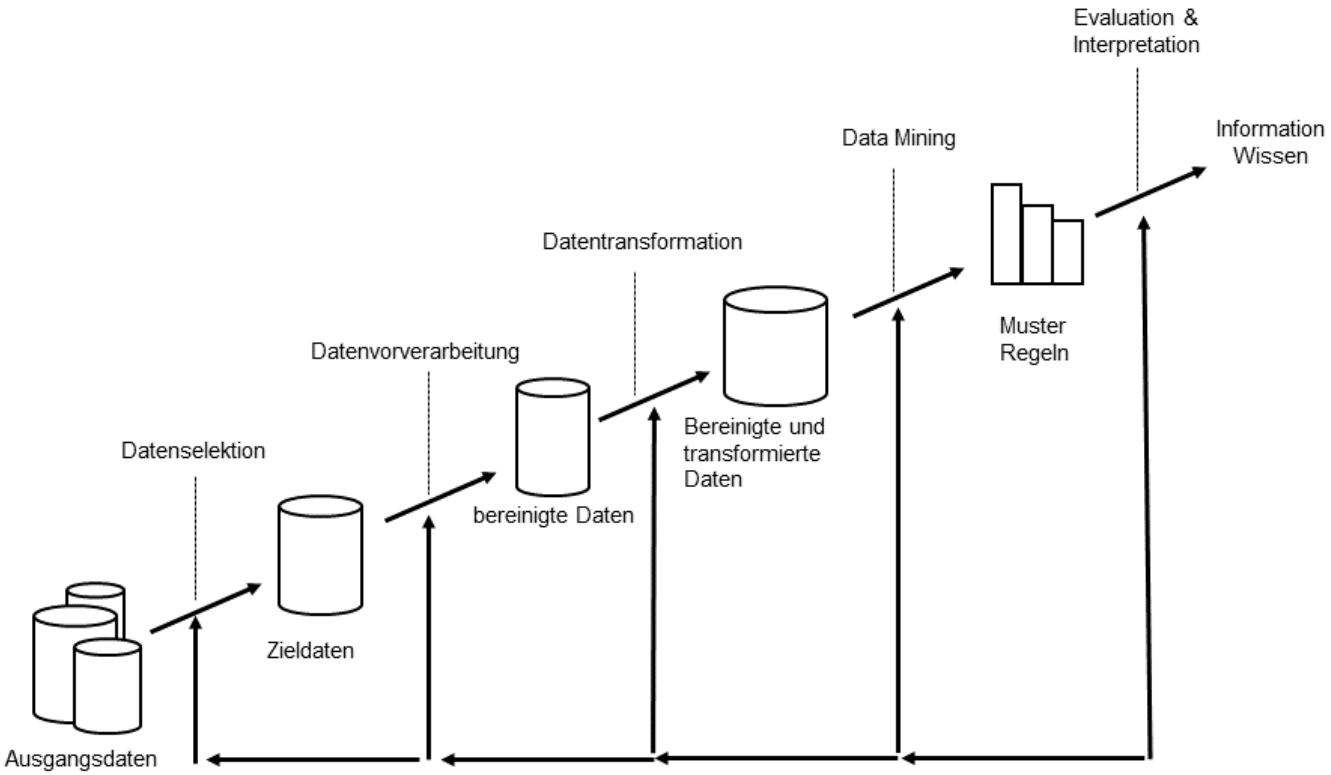


*Abbildung 3 KDD-Stufenmodell nach Fayyad*

Fayyad et al (1996) beschreiben als Ziel ihres Modells die Extraktion von hochwertigem Wissen aus Basisdaten. Einen Überblick der Phasen zur Wissensentdeckung gibt Abbildung 3.

Zu Beginn des Prozesses wird ein Verständnis der Domäne aufgebaut aus dem das KDD-Ziel abgeleitet wird (Fayyad et al. 1996). Diese Phase ist in der Abbildung nicht dargestellt. Durch die Auswahl geeigneter Daten (Datenselektion), sowie der Vorbereitung und Bereinigung des Datensatzes (Datenvorverarbeitung) wird die Datenbasis für folgende Analysen geschaffen, welche durch die Datentransformation an die Anforderungen der ausgewählten Data Mining angepasst wird (Sharafi 2013). Anhand der Data-Mining-Methoden wird nach Mustern gesucht (Data Mining). Die Ergebnisse werden im Abschluss interpretiert, dokumentiert und an die Domänen weitergegeben (Fayyad et al. 1996).

Im Vergleich zum KDD-Prozess, sowie weiteren Modellen, beschreiben (Sharma und Osei-Bryson 2008) das CRISP-DM-Modell als ein detaillierteres Vorgehen, da es besonders die Aspekte zu Beginn eines Projektes abdeckt. Das Modell ist in die sechs Phasen Domänenverständnis, das Datenverständnis, die Datenvorverarbeitung, die Modellierung, die Evaluierung und den Einsatz aufgebaut und in Abbildung 4 dargestellt. (Wirth und Hipp 2000).

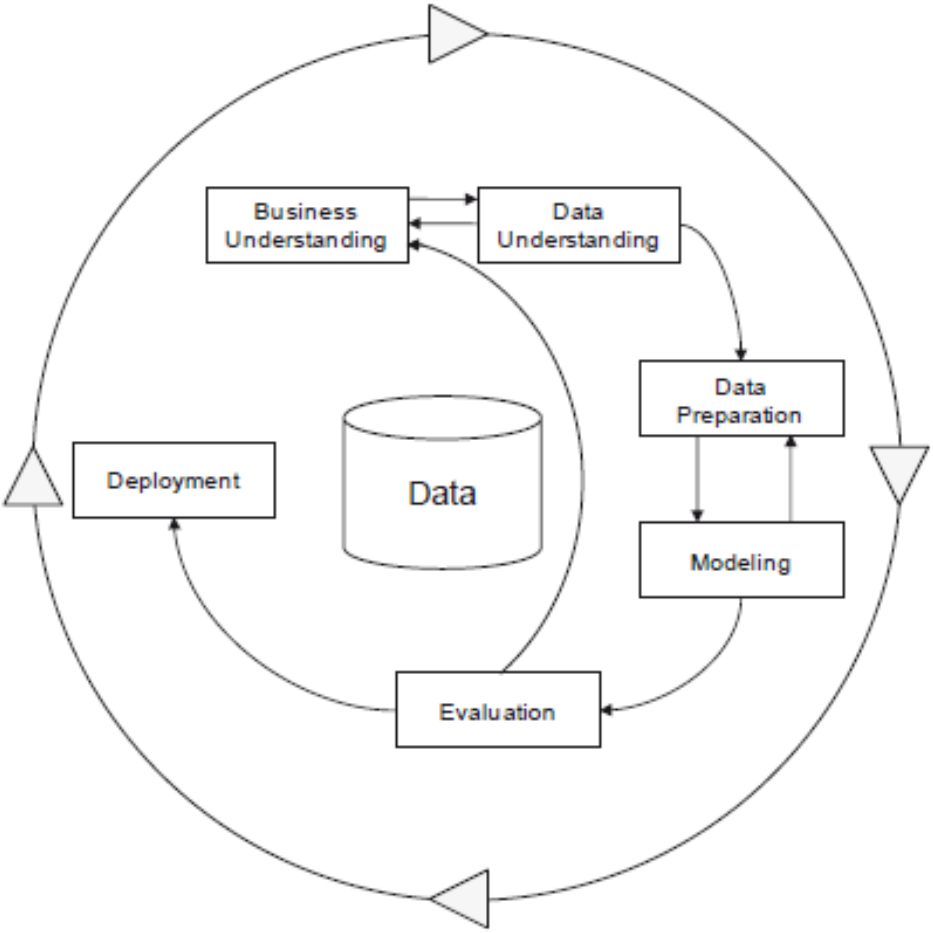


*Abbildung 4 Crisp-DM Modell*

In der Phase des Domänenverständnisses (Business Understanding) wird der Fokus auf die Ziele und die Anforderungen des Projektes gelegt, um diese zu verstehen (Wirth und Hipp 2000). Anhand dieses Wissens wird ein Data Mining Problem definiert und ein Projektplan ausgearbeitet und die definierten Ziele zu erreichen (Wirth und Hipp 2000).

Mit der Formulierung des Data Mining Problems hängt das Verständnis der Daten direkt zusammen (Data Unterstanding), da für die Formulierung ein gewisses Datenverständnis benötigt wird (Wirth und Hipp 2000). In dieser Phase werden Daten gesammelt und bearbeitet, um

erste Einblicke zu Erlangen und etwaige Herausforderungen in der Datenqualität zu erkennen (Wirth und Hipp 2000).

In der Datenvorverarbeitung (Data Preparation) wird der endgültige Datensatz aufgebaut, anhand dessen die späteren Analysen durchgeführt werden (Wirth und Hipp 2000). Zu dieser Phase gehören alle Schritte der Auswahl an Attributen und Tabellen, sowie zur Bereinigung und Transformation des Datensatzess, um ein geeignetes Format für die anschließend genutzte Software bereitzustellen (Wirth und Hipp 2000). Diese Schritte können mehrmals und in unterschiedlicher Reihenfolge durchgeführt werden (Wirth und Hipp 2000).

Anhand der Problemstellung werden in der Modellierungsphase (Modeling) erste Verfahren angewendet (Wirth und Hipp 2000). Die nötigen Parameter werden angepasst, um ein optimales Ergebnis zu erlangen (Wirth und Hipp 2000). Wie in den anderen Phasen auch, hängt die Modellierungsphase eng mit der Datenvorverarbeitung zusammen, da neue Ideen zum Datenaufbau aufkommen oder ein Datenprobleme während der Modellierung auftreten können (Wirth und Hipp 2000).

Zum Zeitpunkt der Evaluation wurden im Projekt verschiedene Modelle erstellt, die aus datenanalytischer Sicht hochwertig erscheinen (Wirth und Hipp 2000). Vor der endgültigen Implementierung des Modells müssen die Modelle gründlich evaluiert und die Schritte zur Modellerstellung überprüft werden (Wirth und Hipp 2000). Ein Hauptziel besteht darin, sicherzustellen, dass keine wichtigen geschäftlichen Aspekte vernachlässigt wurden (Wirth und Hipp 2000). Am Ende dieser Phase sollte eine Entscheidung über die Verwendung der Ergebnisse des Data Minings getroffen werden (Wirth und Hipp 2000).

In der Regel erfordert das erworbene Wissen eine strukturierte Aufbereitung und Präsentation (Einsatz), die für den Kunden nutzbar ist (Wirth und Hipp 2000). Die Phase der Implementierung (Deployment) kann je nach Anforderungen so simpel sein wie die Erstellung eines Berichts oder so komplex wie die Implementierung eines reproduzierbaren Data-Mining-Prozesses (Wirth und Hipp 2000).

Wie aus den oben beschriebenen Vorgehensmodellen herauszulesen ist, wird der Umfang der Datenvorverarbeitung unterschiedlich definiert. Nach Fayyad et al. (1996) ist die Datentransformation eine eigenständige Phase, in welcher der Datensatz spezifisch auf die folgenden Data Mining Verfahren angepasst wird. Nach (Wirth und Hipp 2000) oder (Runkler 2020) ist dieser Arbeitsschritt jedoch Teil der Datenvorverarbeitung. Um ein besseres Verständnis der Datenvorverarbeitung zu erhalten, wird diese im folgenden Kapitel mit Ausnahme der Datentransformation näher betrachtet.

### 2.1.3 Datenvorverarbeitung

Umfangreiche Datensätze erfordern eine sorgfältige Vorbereitung und Anpassung, bevor fortgeschrittene Analysemechanismen wie Data Mining angewendet werden können (Luengo et al. 2020). Rohdaten sind anfällig für fehlende und inkonsistente Daten, Rauschen und Ausreißern (Alasadi und Bhaya 2017). Die Ursache für fehlende Werte können Werte sein, die außerhalb eines Messbereiches liegen oder durch unterschiedliche Messzeitpunkte von Sensoren aufgenommen wurden (Mishra et al. 2020). Rauschen und Ausreißer können nach Runkler (2020) in systematische und zufällige Fehler kategorisiert werden. Zufällige Fehler sind beispielsweise Mess- oder Übertragungsfehler, welche als Ausreißer oder additives Rauschen modelliert werden (Runkler 2020).

Rauschen beschreibt zufällig aufgezeichnete Fehler, durch die der Wert einer Variable verändert wird (Aggarwal 2015). Allerdings kann ein aufgezeichneter Fehler in dem vorliegenden Datensatz seine Gültigkeit besitzen (Bramer 2020). García et al. (2015) unterscheidet Rauschen in Klassen- und Attributrauschen. Beim Klassenrauschen handelt sich um falsch beschriftete Daten, die durch Eingabefehler oder unzureichende Information entstehen. Dabei kann es sich um Daten handeln, die sich im Datensatz doppeln, aber eine unterschiedliche Klassenbezeichnung besitzen (widersprüchliche Daten), oder deren Klassifizierung sich von

der wahren Bezeichnung unterscheidet (Fehlklassifizierung) (García et al. 2015). Das Attributrauschen beschreibt verfälschte Attributwerte, die unvollständige Attribute enthalten oder bei denen Attributwerte fehlen oder unbekannt sind (García et al. 2015).

Ausreißer sind Datenpunkte, die stark von den anderen Punkten abweichen (Baumann et al. 2018). Ausreißer entstehen beispielsweise durch falsche Verarbeitung oder Erfassung der Daten (Runkler 2020). Diese Werte werden oft als Rauschen markiert und daraufhin gelöscht (Cleve und Lämmel 2020). Allerdings können die Werte wie beim Rauschen korrekt erfasste Werte sein (Cleve und Lämmel 2020). Aggarwal (2017) bezeichnet diese Fehler als Abnormalitäten, Abweichungen oder widersprüchliche Beobachtungen. Ausreißer können neben der Verschlechterung der Datenqualität Informationen erhalten, aus denen Rückschlüsse auf den Prozess der Datenerzeugung gewonnen werden können (Aggarwal 2017). Daher ist die Wahl der richtigen Datenvorverarbeitungsschritte von hoher Bedeutung.

Eine schlechte Qualität der Daten hat Einfluss auf die späteren Ergebnisse, weswegen die Daten durch bearbeitet werden müssen, um so die Grundlage für aussagekräftige Ergebnisse zu bilden (Alasadi und Bhaya 2017). Das Ziel der Datenvorverarbeitung besteht darin, einen Datensatz zu erstellen, der als zuverlässiger und geeigneter Input für die Anwendung von Data Mining dient (Luengo et al. 2020). Zu diesem Zweck ist die Datenvorverarbeitung in die vier Schritte Datenintegration, Datenbereinigung, Datentransformation und Datenreduktion eingeteilt (Tamilselvi et al. 2015).

Während der Datenintegration werden Daten aus verschiedenen Quellen in einem Speicherort miteinander verbunden (Cleve und Lämmel 2020). Quellen können mehrere Datenbanken oder Dateien sein, in denen die Daten gespeichert werden (Tamilselvi et al. 2015). Bei fehlerhafter Ausführung der Datenintegration kann es zu Redundanzen oder Inkonsistenzen im Datensatz kommen (García et al. 2015).

Die Datenbereinigung hat das Ziel, fehlende Werte auszufüllen, Rauschen zu behandeln, Ausreißer zu erkennen und Inkonsistenzen in den Daten zu korrigieren (Tamilselvi et al. 2015). Um fehlende Werte zu füllen, gibt es mehrere Verfahren. Fehlende Tupel werden als eine Möglichkeit ignoriert (Tamilselvi et al. 2015). Allerdings ist diese Herangehensweise sehr von Nachteil, wenn die Anzahl der fehlenden Werte pro Attribut stark schwankt (Tamilselvi et al. 2015). Sollen die Werte gefüllt werden, kann dies durch manuelles Einfügen von Werten oder einer globalen Konstante geschehen, wobei die manuelle Bearbeitung bei großen Datensätzen zu einem erheblichen Zeitaufwand führt (Tamilselvi et al. 2015). Durch einfügen einer Konstante für alle fehlenden Werte, kann das spätere angewendete Mining-Verfahren möglicherweise ein Muster erkennen (Tamilselvi et al. 2015). Neben einer einheitlichen Konstante, besteht die Herangehensweise, fehlende Werte anhand der Bedeutung des Attributs oder der Zugehörigkeit zu einer bestimmen Kategorie zu füllen (Tamilselvi et al. 2015). Beispielsweise wird für die Produktion von einem Produkt der Wert von 3000 Einheiten angenommen,Als weiteres Beispiel wird einem Kunden eine bestimmte Kategorie zugeordnet, wird für den Vertrieb eines Produktes der Durchschnitt aller Kunden in derselben Kategorie eingetragen. Um Rauschen zu behandeln werden die Verfahren Binning, Regression und Cluster-Analyse genutzt (Cleve und Lämmel 2020). Beim Binning werden Werte mit Werten in ihrer Umgebung zu sogenannten Bins zusammengefasst (Tamilselvi et al. 2015). Die Werte innerhalb der Bins werden daraufhin durch Mittel- oder Grenzwerte ersetzt, was einen glättenden Effekt zur Folge hat (Cleve und Lämmel 2020). Bei der Regression werden Werte durch eine Funktion ersetzt (Cleve und Lämmel 2020). Bei einer linearen Regression wird die „beste“ Gerade zwischen zwei Werten oder mehreren gesucht, sodass ein Wert aus einem anderen Vorhergesagt werden kann (Tamilselvi et al. 2015). Bei der Clusteranalyse werden ähnliche Ausreißer gruppiert und von den anderen Daten getrennt (Cleve und Lämmel 2020).

Durch die Datenreduktion soll das Volumen des Datensatzes unter Berücksichtigung der Aussagekraft verkleinert werden (Alasadi und Bhaya 2017). Die Größe eines Datensatzes kann die Leistung der Data Mining-Verfahren stark beeinflussen (Petersohn 2009). Das Volumen kann bei der Datenreduktion durch die Aggregation der Attribute oder dem Entfernen von unwichtigen Attributen erfolgen (Tamilselvi et al. 2015). Zudem kann eine Dimensionsreduktion

durch Verkleinerung der Datensätze erreicht werden (Petersohn 2009). Durch Datendiskretisierung werden Rohdatenwerte durch Wertebereiche oder höhere abstrakte Ebenen ersetzt (Tamilselvi et al. 2015). Diskretisierung und die Generierung von Konzepthierarchien sind mächtige Werkzeuge für das Data Mining, da sie das Mining von Daten auf mehreren Abstraktionsebenen ermöglichen (Tamilselvi et al. 2015).

### 2.1.4 Datentransformation

In dem vorangegangenen Kapitel wurden Herausforderungen betrachtet, welche unabhängig, von den angewendeten Analyseverfahren durchgeführt werden können. Allerdings sind Daten in ihrer Ausgangsform nicht für die gewünschten Analyseverfahren geeignet (Cleve und Lämmel 2020). Mittels Datentransformation sollen die Daten in eine Form umgewandelt werden, welche für die angewendeten Analysemethoden geeignet ist (Schulz et al. 2022). In der Literatur bestehen verschiedene Ansichten, welche Schritte der Datentransformation zugeordnet werden. Wie in Kapitel 2.1.3 dargestellt, sehen Fayyad et al. (2020) die Datentransformation als eine eigene Phase an, in welcher nützliche Funktionen ausgewählt werden, um die Daten entsprechend des gewünschten Ziels ausgewählt werden (Fayyad et al. 1996). Entsprechend des Zieles, werden Attribute festgelegt, an welche die Daten so angeglichen werden (Sharafi 2013). Bei dem CRISP-DM Modell ist die Datentransformation ein Schritt in der Datenvorverarbeitung zur Erstellung eines geeigneten Formats (Sharafi 2013). Je nach Autor und Betrachtung des Begriffs Datentransformation werden dieser unterschiedlichen Verfahren zugeordnet. Daher betrachtet diese Arbeit in diesem Kapitel die Schritte, in welchen die Datentypen oder die Form des Datenbestandes verändert werden, und stellt diese im Folgenden vor.

#### Anpassung von Datentypen und Konvertierungen

Bei Datenbeständen, wie relationalen Datenbanken, kann ein Attribut sowohl als integer als auch Character dargestellt werden (Cleve und Lämmel 2020). In diesem Fall muss geklärt werden, ob der Datentyp umgewandelt werden muss (Cleve und Lämmel 2020).

Eine **Anpassung von Zeichenketten** muss gegebenenfalls angewendet werden, um Umlaute, Groß- und Kleinschreibung oder Leerzeichen umzuwandeln (Cleve und Lämmel 2020).

#### Anpassung von Datumsangaben und Maßeinheiten

Datentypen besitzen aufgrund von unterschiedlichen Formaten in verschiedenen Ländern eine variierende Codierung, oder können in unterschiedlichen Zeitzonen erfasst worden sein und benötigend daher eine Vereinheitlichung (Cleve und Lämmel 2020). Die Anpassung der Datumsangaben, Maßeinheiten oder der Zeichenketten muss nach Cleve und Lämmel (2020) gegebenenfalls während der Datenintegration durchgeführt werden.

#### Kombination oder Separierung von Attributen

Es kann im Datensatz nötig sein, verschiedene Attribute voneinander zu separieren (Cleve und Lämmel 2020). Beispielsweise kann das Attribut *Name* in die Attribute *Vorname* und *Nachname* getrennt werden. Im Gegensatz dazu kann es erforderlich sein, die Attribute *Tag, Monat* und *Jahr* zu einem neuen Attribut *Datum* zusammenzuführen.

#### Berechnung abgeleiteter Werte

Durch die Berechnung von abgeleiteten Werten können neue Attribute eingeführt werden (Schulz et al. 2022). Das Attribut *Gewinn* wird beispielsweise als Differenz zwischen den Attributwerten *Ausgaben* und *Einnahmen* ermittelt (Cleve und Lämmel 2020). Das Alter wiederum kann als Differenz zwischen dem aktuellen Datum und dem Geburtsdatum bestimmt werden (Cleve und Lämmel 2020). Durch die Ableitungen können zudem Typkonvertierungen vorgenommen werden, um beispielsweise aus Städtenamen die jeweiligen Postleitzahlen, oder umgekehrt, abzuleiten (Schulz et al. 2022).

**Datenaggregation**

Die Datenaggregation wurde bereits unter der Datenreduktion genannt. Allerdings kann dieses Verfahren genutzt werden, um neue Daten zu erhalten (Cleve und Lämmel 2020). In vielen Fällen liegen Attribute in einer zu feinen Aggregationsebene vor (Cleve und Lämmel 2020). Bei der Aggregation werden verschiedene Attributwerte zu einem zusammengefasst, sodass eine gröbere Aggregationsebene erreicht wird (Cleve und Lämmel 2020).

**Generalisierung**

In diesem Kontext erfolgt die Transformation, wobei niedrigstufige Daten mittels Konzept-Hierarchien durch höherstufige Daten ersetzt werden (Tamilselvi et al. 2015). Beispielsweise können Attribute, wie Straßen, durch abstraktere Attribute, wie beispielsweise Städte, auf einer höheren Ebene ersetzt werden.

**Datenglättung**

Durch Datenglättung wird jeder numerische Wert durch einen idealisierten Wert zu ersetzt (Cleve und Lämmel 2020). Dieser Ansatz wurde bereits in Kapitel 2.1.3 in Form der Datenbereinigung beschrieben, um Rauschen und Ausreißer zu bearbeiten. Das Ziel in diesem Schritt ist die ursprüngliche Wertemenge zu reduzieren, um so bessere Lösungen zu erhalten (Cleve und Lämmel 2020). Zu diesen Techniken gehören nach Baskar et al (2013) die bereits beschriebenen Verfahren:

- Binning
- Clustering
- Regression

**Umwandlung von nominalen und ordinalen Daten in metrische Daten**

Liegt ein Attribut mit ordinaler Ausprägung vor, so kann jeder Gruppe ein Zahlenwert zugeordnet werden (Cleve und Lämmel 2020). Dies kann direkt in einer normierten Weise umgesetzt werden. Beispielsweise kann einem Attribut Geschwindigkeit mit den Werten sehr schnell, mittelschnell und langsam die Werte 1, 0,5 und 0 zugeordnet werden (Cleve und Lämmel 2020). Es muss jedoch darauf geachtet werden, dass die Ordnungsrelation eingehalten wird (Cleve und Lämmel 2020). Die Binärcodierung von Attributen oder Binning sind weitere Möglichkeiten Daten umzuwandeln (Cleve und Lämmel 2020).

**Diskretisierung**

Bei diesem Verfahren werden Zahlenwerte durch Kategorien ersetzt (Han et al. 2012). Beispielsweise kann das Gewicht eines Produktes in die Kategorien schwer und leicht eingeteilt werden. Auf diese Weise werden Informationen zu Gruppen zusammengefasst und so die Attributsausprägung reduziert. Zudem erfolgt durch die Diskretisierung eine Art Filterung von verrauschten Daten.(Schulz et al. 2022).

**Normalisierung, Skalierung und Standarisierung**

Normalisierung ist eine Art der Skalierung (Cleve und Lämmel 2020). Bei der Skalierung, werden alle Werte auf einen Bereich skaliert (Cleve und Lämmel 2020). Bei der Normalisierung ist dies ein bestimmter Wertebereich, auf den die Daten skaliert werden (Cleve und Lämmel 2020). Während bei der Normierung die numerischen Werte auf ein bestimmtes Intervall abgebildet werden, sorgt die Standardisierung dafür, dass die Werte um den Mittelwert Null mit einer gegebenen Standardabweichung von Eins streuen (Schulz et al. 2022).

## 2.2 Produktion und Logistik

Nachdem im vorangegangenen Kapitel die Grundlage für die spätere Literaturrecherche gelegt wurde, betrachtet dieses Unterkapitel die Domäne Produktion und Logistik. In den beiden Kapiteln wird zunächst auf die Produktion als Wertschöpfungsprozess eingegangen, um daraufhin auf die Grundlagen der Logistik einzugehen.

### 2.2.1 Produktion als Wertschöpfungsprozess

Unternehmen sind organisatorische Einheiten, die das Sachziel verfolgen, Güter und Dienstleistungen zu erstellen (Grün und Jammernegg 2019). Der Begriff der Produktion wird daher in der Literatur in einem weiten und einem engen Sinn abgegrenzt (Küpper und Helber 2004). In der weiten Fassung bezieht sich der Begriff auf jede Erstellung von Sachgütern und Dienstleistungen (Küpper und Helber 2004). Hierzu gehören Forschungs- und Entwicklungsleistungen, die Beschaffung der Einsatzgüter oder der Absatz der Produkte (Küpper und Helber 2004). Dagegen wird Produktion im engeren Sinne als Synonym zu Fertigung verwendet (Küpper und Helber 2004). Diese Betrachtung bezieht sich nur auf die Erstellung von materiellen (Sach-) Gütern oder immateriellen Dienstleistungen, wie menschlicher oder maschineller Arbeit (Küpper und Helber 2004). Vereinfachend wird unter dem Begriff Produktion die „Erzeugung von Ausbringungsgütern (Produkten) aus materiellen und nichtmateriellen Einsatzgütern (Produktionsfaktoren) nach bestimmten technischen Verfahrensweisen“ (Günther und Tempelmeier 2013, S. 6) verstanden.

Dieser Produktionsprozess besteht aus einzelnen Abschnitten, welche einen Teilprozess der Produktion umfassen und in geeigneten organisatorischen Einheiten zusammengefasst sind (Günther und Tempelmeier 2013). Die organisatorischen Einheiten werden als Arbeitssystem bezeichnet, welches in Abbildung 5 nach Günther und Tempelmeier (2013) dargestellt ist:

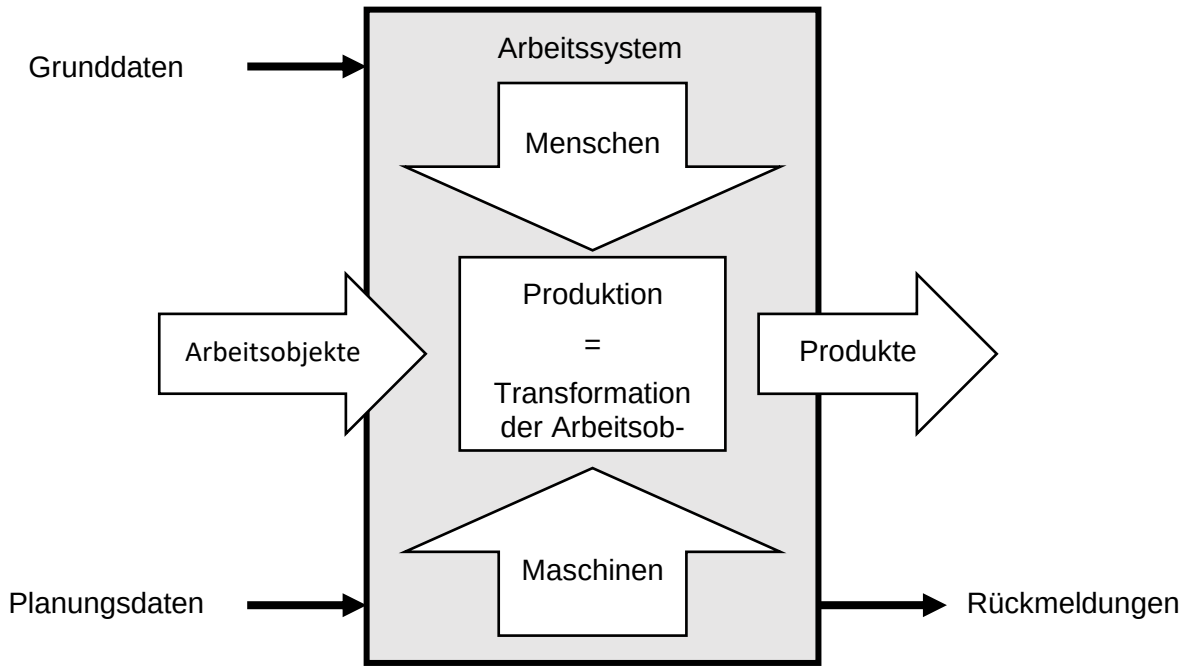


*Abbildung 5 Aufbau eine Arbeitssystems*

Das Arbeitssystem ist die kleinste eigenständig arbeitende Einheit in einem Produktionssystem und wird durch die Elemente Input, Transformation und Output beschrieben (Günther und Tempelmeier 2013).

**Input-Einheiten** sind die zu bearbeitende Vorprodukte (Arbeitsobjekte). Aus den Grunddaten werden Informationen erschlossen, welche für die Produktion benötigt werden (Günther und

Tempelmeier 2013). Die Planungsdaten geben beispielsweise Informationen zu Anzahl der Erzeugnisse oder dem Zeitpunkt der Fertigstellung (Günther und Tempelmeier 2013).

**Transformation** beschreibt den eigentlichen Produktionsvorgang. In diesem Prozess erfolgt die Umwandlung der Arbeitsobjekte in Produkte durch Statusänderung und Wertsteigerung der Objekte (Günther und Tempelmeier 2013).

**Output-Einheiten** sind die im Transformationsprozess erzeugten Produkte. Diese können Güter oder Dienstleistungen sein (Günther und Tempelmeier 2013). Der Zeitpunkt der Fertigstellung wird an Produktionsplanungs- und Steuerungssystem zurückgemeldet (Günther und Tempelmeier 2013).

In einer Wettbewerbswirtschaft, verfolgt der Produzent das Ziel, Leistungen mit einer möglichst hohen Wertschöpfung (Wertsteigerung) am Markt zur Verfügung zu stellen (Grün und Jammernegg 2019). Daher ist die Produktion ein Wertschöpfungsprozess, in dem aus einfachen oder komplexen Gütern, wie Material oder Daten, wertgesteigerte Output-Einheiten erzeugt werden (Grün und Jammernegg 2019). Nach Grün und Jammernegg (2019) müssen folgende Anforderungen erfüllt werden, um eine Wertschöpfung zu erzielen

**Zeit**

Die Erzeugung eines Produkts erfordert eine Vielzahl von Schritten, die jeweils eine bestimmte Zeit zu ihrer Ausführung benötigen (Günther und Tempelmeier 2013). Je schneller diese zeitliche Strecke überwunden wird, umso größer ist die Wertschöpfung (Günther und Tempelmeier 2013). Dieses Ziel spiegelt in dem Bestreben nach kurzen Durchlaufzeiten wider (Günther und Tempelmeier 2013).

**Flexibilität**

Flexible Produktionsprozesse besitzen die Fähigkeit, sich an veränderte Umweltbedingungen, wie der Änderung des Produktprogramms anpassen zu können (Grün und Jammernegg 2019). Zur Charakterisierung der Flexibilität, ist der Anpassungsumfang, die wirtschaftlichen Auswirkungen und die Zeit von Bedeutung. Aus strategischer Sicht gilt ein Produktionssystem als flexibel, wenn es sich in angemessener Zeit auf die veränderten Umweltbedingungen einstellen kann (Günther und Tempelmeier 2013). Aus operativer Sicht ist die Fähigkeit von Bedeutung, kurzfristig auf notwendige Veränderungen zu reagieren (Günther und Tempelmeier 2013).

**Qualität**

Die Leistung eines Produktionssystems lässt sich neben dem mengen- und wertmäßigen Output, auch in der Qualität und die daraus resultierende Kundenzufriedenheit messen (Günther und Tempelmeier 2013). Hierzu gehören beispielsweise Funktionalität oder der Zuverlässigkeit der erzeugenten Produkte (Günther und Tempelmeier 2013).

**Wirtschaftlichkeit**

Die Wirtschaftlichkeit lässt sich durch die Formulierungen des Maximum-, sowie des Minimumprinzips beschreiben (Günther und Tempelmeier 2013). Bei dem Maximumprinzip ist mit einem gegebenen Inputgütern ein maximales Ergebnis zu erreichen (Günther und Tempelmeier 2013). Nach dem Minimumprinzip, sollen mit einem minimalem Wert an Inputgütern ein vorgegebenes Ergebnis erreicht werden (Günther und Tempelmeier 2013).

In den meisten Unternehmen sind die Arbeitssysteme miteinander verknüpft, wodurch sich ein Materialfluss ergibt, weswegen die Arbeitssysteme durch Transportwege miteinander verknüpft werden müssen (Günther und Tempelmeier 2013). Zudem werden die benötigten Ressourcen extern beschafft (Günther und Tempelmeier 2013). Die Wirtschaftlichkeit eines Produktionssystems hängt außerdem von der Infrastruktur ab, in welcher die Produktion stattfindet (Günther und Tempelmeier 2013). Während die Produktion Prozesse zur Güter- und Dienstleistungserstellung betrachtet, werden die zuvor genannten Punkte von der Logistik abgedeckt., da diese eine Querschnittsfunktion besitzt und sich über alle Phasen erstreckt (Küpper und Helber 2004).

### 2.2.2 Grundlagen Logistik

Für den Begriff der Logistik gibt es mehrere Definitionen (Pfohl 2018). Die Logistik hat die Pflicht, dass „ein Empfangspunkt gemäß seines Bedarfs von einem Lieferpunkt mit dem *richtigen Produkt* (in Menge und Sorte), im *richtigen Zustand,* zur *richtigen Zeit,* am *richtigen Ort* zu den dafür minimalen Kosten versorgt wird." (Pfohl 2018, S. 12). Dabei agiert die Logistik funktions- und unternehmensübergreifend, wobei sie das Gesamtbild betrachtet und sich an Nutzen und Service orientiert (Martin 2016). Wird die flussorientierte Definition der amerikanischen Logistikgesellschaft Council of Supply Chain Management Professionals (CSCMP) betrachtet, so definiert diese Logistik als den „Prozess der Planung, Realisierung und Kontrolle des effizienten, kosteneffektiven Fließens und Lagerns von Rohstoffen, Halbfabrikaten und Fertigfabrikaten und den damit zusammenhängenden Informationen vom Lieferzum Empfangspunkt entsprechend den Anforderungen des Kunden." (Pfohl 2018, S. 12)

**Transportprozesse** sind für die räumliche Verschiebung der Güter im Logistikprozess verantwortlich (Huber und Laverentz 2019).

**Umschlagsprozesse** beziehen sich auf Aktivitäten wie bei Transport- oder Ladehilfsmittel eines Gutes gewechselt werden sollen (Huber und Laverentz 2019) Hierbei werden zum Beispiel Güterströme zusammengefasst, umsortiert oder auf ausgehende Güterströme verteilt (Huber und Laverentz 2019).

In den **Lagerprozessen** erfolgt eine zeitliche Transformation der Güter, indem sie beispielsweise zwischen ihrer Bereitstellung und Verwendung gelagert werden (Huber und Laverentz 2019).

**Bearbeitungsprozesse** verändern die Güter in ihrer physischen Form (Huber und Laverentz 2019). Diese Schritte können in physischen Transformation (Verpacken und Etikettieren) und physisch neutralen Bearbeitungsprozessen wie Prüfabwicklungen unterschieden werden (Huber und Laverentz 2019).

**Informationsprozesse** haben keine unmittelbare Auswirkung auf die Güter selbst, sondern betreffen die Transformation von Informationsobjekten, wie beispielsweise Material- und Ladehilfsmitteldaten (Huber und Laverentz 2019). Um die Güter eindeutig identifizieren zu können, werden die Informationen im Lagerverwaltungssystem mit den Gütern verknüpft (Huber und Laverentz 2019).

Um einen effizienten und kosteneffektiven Güterfluss sicherzustellen, müssen diese Prozesse in verschiedenen Unternehmensbereichen nahtlos koordiniert werden (Pfohl 2018). Dies bedeutet, dass Logistik als eine querschnittliche Funktion innerhalb des Unternehmens betrachtet werden kann und sich über die drei grundlegenden betrieblichen Funktionen - Beschaffung, Produktion und Absatz - erstreckt (Pfohl 2018). Nach Oeldorf und Olfert (2018) teilt sich die Logistik daher in die vier Teilbereiche Beschaffungs-, Produktions-, Distributions- und Entsorgungslogistik auf.

Wichtige logistische Kennzahlen sind hierbei nach Martin (2016) beispielsweise

- Servicegrad
- Auftragsdurchlaufzeiten
- Entsorgungskosten
- Flächennutzungsgrad
- Verfügbarkeiten von Materialien

# 3 Untersuchung von Datentransformation

Im vorangegangenen Kapitel wurde mit der Betrachtung der Datentransformation im Kontext der Datenvorverarbeitung der Grundstein für die systematische Literaturrecherche gelegt. Diese wird im Kapitel 3.1 näher dargestellt und daraufhin angewendet. Im Kapitel 3.2 wird das Ergebnis dargestellt. Hierzu wird ein exemplarischer Vergleich zwischen den Publikationen durchgeführt, in welchem die dargestellten Verfahren gegenübergestellt werden. Daraus folgt in Kapitel 3.3 die Darstellung der Datentransformation mit einer genaueren Vorstellung der Verfahren.

## 3.1 Methodik der systematischen Literaturrecherche

Literaturübersichten spielen in der Wissenschaft eine große Rolle (vom Brocke et al. 2009). Dies begründet vom Brocke et al. (2009) darin, dass durch die Kombination und Interpretation von bestehendem Wissen, neues Wissen erlangt wird. Eine Literaturübersicht ist nach vom Brocke et al. (2009) eine Zusammenfassung in einer Domäne, welche dabei unterstützt spezifische Recherchefragen abzuleiten. Dabei hat die Qualität der Literatursuche einen hohen Einfluss auf die Qualität der Literaturübersicht (vom Brocke et al. 2009). In dieser Bachelorarbeit soll die Datentransformation im Kontext der Datenvorverarbeitung in Produktion und Logistik auf der Grundlage bereits bestehender Verfahren vorgestellt werden. Daher wird eine systematische Literaturrecherche durchgeführt, um diese verstehen zu können. Dieses Konzept wird im Folgenden nach dem bekannten System von vom Brocke et al. (2009) dargestellt. Darauf aufbauend wird das Vorgehen der in dieser Arbeit umgesetzten Literaturrecherche beschrieben.

Wie in der Einleitung angedeutet, ist die Literaturrecherche die Grundlage, bestehendes Wissen zusammenzufassen und durch Erkennen das von Verknüpfungen neues Wissen zu generieren (vom Brocke et al. 2009). Durch systematische Literaturrecherchen wird die Relevanz und Strenge der Recherchen beibehalten, da zum einen die Untersuchung von bereits bekanntem Wissen vermieden und bestehendes Wissen effektiv genutzt wird. Um Wissen zu erlangen müssen qualitativ hochwertige Quellen identifiziert werden (vom Brocke et al. 2009). Daher müssen während der Literaturrecherche diese Quellen identifiziert und diese darauf zu prüfen, in wie weit sie auf die Studie angewendet werden können (vom Brocke et al. 2009).


*Abbildung 6 Aufbau strukturierte Literaturrecherche*

Die systematische Literaturrecherche besteht nach vom Brocke et al. (2009), wie in Abbildung 6 zu sehen ist, aus den fünf Phasen Definition des Rechercheumfangs, Konzepterstellung, Literaturrecherche, Literaturanalyse und der Forschungsagenda. Der Umfang der Recherche wird in der ersten Phase „Definition des Rechercheumfangs“ festgelegt. Vom Brocke et al. (2009) nennen hierzu die Taxonomie für Literaturrecherchen nach Cooper (1988). Des Weiteren wird in dieser Phase die Forschungsfrage definiert (vom Brocke et al. 2009).

*Tabelle 1 Taxonomie nach Cooper*

| Merkmal | Kategorie |
|---|---|
| Fokus | Forschungsergebnisse, Forschungsmethoden, Theorie, Praktiken, Anwendungen |
| Ziele | Integration, Kritik, Identifikation zentraler Aspekte |
| Organisation | Historisch, methodisch, konzeptionell |
| Perspektive | Neutrale Darstellung, Vertretung eines Standpunktes |
| Zielgruppe | Spezialisierte Wissenschaftler, allgemeine Wissenschaftler, öffentliches Publikum |
| Abdeckungsgrad der Quellen | Vollständig, vollständig selektiv, repräsentativ, zentrale Abdeckung |

Die Taxonomie nach Cooper besteht, wie aus Tabelle 1 erkennbar, aus den sechs Merkmalen des Fokus, der Ziele, der Organisation, der Art der Darstellung, an wen die Arbeit gerichtet sein soll und wie weit die vorhandene Literatur abgedeckt werden soll.

Um den Fokus der Literaturrecherche zu wahren, wird durch dieses Merkmal festgelegt, was in der Recherche dargestellt werden soll (Cooper 1988). Durch das zweite Merkmal soll der Verfasser das Ziel der Ausarbeitung festlegen (Cooper 1988). Dies kann die Integration bzw. Synthese früherer Literatur sein, welche vom Autor als für das gleiche Thema relevant angesehen werden (Cooper 1988). Allerdings können ebenso eine kritische Analyse bestehender Literatur, die Identifikation zentraler Probleme und Fragestellungen Ziele der Ausarbeitung sein (Cooper 1988). Die Organisation der Recherche legt fest, wie die Übersicht angeordnet wird. Dies kann historisch, also in einer chronologischen Reihenfolge der Publikationen, methodisch, also dass Publikationen mit derselben Abstrakten Idee zusammen dargestellt werden, oder konzeptionell, also die Gruppierung von ähnlichen Methoden in Unterkategorien, sein (Cooper 1988). Die Perspektive legt fest, wie die Ansicht des Verfasser die Übersicht beeinflusst (Cooper 1988). Die Zielgruppe legt fest, an wen sich die Übersicht richtet (Cooper 1988). Diese Einordnung beeinflusst vor allem den Schreibstil (Cooper 1988). Der Abdeckungsgrad als sechstes Merkmal ist laut Cooper (1988) vermutlich der relevanteste Aspekt einer Literaturübersicht. Durch vollständige Abdeckung zielt der Verfasser darauf ab, sämtliche Literatur zu erfassen, die existiert (Cooper 1988). Vollständige selektive Abdeckung zieht ebenfalls Schlussfolgerungen aus der gesamten Literatur heran, jedoch werden ausgewählte Beispiele näher beschrieben (Cooper 1988). Durch repräsentative Abdeckung wird die Literatur dargestellt, welche sich im gesamten Spektrum wiederspiegelt (Cooper 1988). Anhand des letzten Abdeckungsgrades sollen Veröffentlichungen dargestellt werden, welche zentral für ein Thema sind (Cooper 1988). Die Durchführung dieser Taxonomie beeinflusst den in Phase drei durchgeführten Suchprozess.

Bei der "Konzepterstellung" werden die zur Suche verwendeten Datenbanken, sowie die Schlüsselbegriffe und Ein- und Ausschlusskriterien definiert (Vom Brocke et al. 2009). Hierzu muss zunächst ein Überblick über bestehendes Wissen geschaffen werden, um die Suchstrategie abzuleiten (vom Brocke et al. 2009).

Während der Literaturrecherche als dritte Phase wird die eigentliche Suche durchgeführt (vom Brocke et al. 2009). Hierzu werden die vordefinierten Suchbegriffe in den festgelegten Datenbanken eingegeben (vom Brocke et al. 2009). Falls nötig, kann die Anzahl der Resultate durch die Operatoren AND oder OR eingeschränkt werden. Vom Brocke betont, dass die verwendeten Suchbegriffe genau dokumentiert werden sollen (vom Brocke et al. 2009). Anhand der

Dokumentation können dritte Nachvollziehen, ob die Auswahl der Suchbegriffe zu der Forschungsfrage passt (vom Brocke et al. 2009). Zudem kann durch Vor- und Rückwärtssuche weitere Literatur gefunden werden (vom Brocke et al. 2009).

In der "Literaturanalyse" als vierte Phase werden die gesammelten Quellen anhand der Titel Abstracts oder des gesamten Textes analysiert (vom Brocke et al. 2009). Um die Literatur zu analysieren, schlagen vom Brocke et al. (2009) eine Konzeptmatrix vor, in welcher die Literatur unterschiedlichen Analyseeinheiten zugeordnet wird und so bestehende Forschungsergebnisse besser arrangiere, diskutieren oder synthetisieren zu können

In der fünften Phase soll die Synthese der Literatur in der "Forschungsagenda" münden(vom Brocke et al. 2009), Diese Agenda besteht aus präziseren und erkenntnisreichen Fragen für künftige Forschungen (vom Brocke et al. 2009)

In Bezug auf den Forschungsumfang dieser strukturierten Literaturrecherche, wie mithilfe der Taxonomie von Cooper (1988) festgelegt, liegt der Fokus auf der Vorstellung der Datentransformation im Kontext der Datenvorverarbeitung in der Knowledge Discovery. Das Hauptziel dieser strukturierten Literaturrecherche besteht darin, einen Überblick über bestehende Behandlungsverfahren zur Datentransformation in der Datenvorverarbeitung für Data Mining zu geben und zentrale Aspekte zu identifizieren. Die Ergebnisse werden neutral präsentiert und richten sich an ein breites wissenschaftliches Publikum. Zudem wurde eine repräsentative Abdeckung von Quellen ausgewählt.

Um sich mit dem Thema vertraut zu machen und das von vom Brocke et al. angesprochene Wissen aufzubauen, wurden anhand des Titels der Arbeit Schlagwörter wie Produktion und Logistik abgeleitet und in der Datenbank der Universitätsbibliothek der TU Dortmund eingegeben. Zudem wurden bereits bearbeitete Abschlussarbeiten herausgesucht, um ein Verständnis der Knowledge Discovery zu erhalten. Daraus resultierend ergaben sich die ersten Quellen, welche für den Stand der Technik genutzt wurden.

In der Phase der Konzepterstellung wurden die Datenbanken für die Quellensuche definiert, darunter ACM, Scopus, Springer Link, IEEE Xplore Digital Library. Es wurden Publikationen berücksichtigt, die zwischen 1997 und 2023 veröffentlicht wurden, da das Modell von Fayyad im Jahr 1996 erstmals in der Literatur erwähnt wurde. Pro Suche wurden die ersten 100 Suchergebnisse betrachtet, welche nach Relevanz sortiert wurden. Zu Beginn der Rechereche wurde die Begriffe Data Mining, Datenvorverarbeitung und Datentransformation genutzt. Allerdings wurden die Begriffe in die Begriffe „Data preprocessing“, sowie „Data transformation“, übersetzt, da englische Datenbanken genutzt wurden.

In der dritten Phase der Literaturrecherche wurde zunächst der Begriff „data transformation“ in dem Suchfeld eingegeben und die Suche auf den Titel beschränkt. Zudem wurden Anführungszeichen genutzt, sodass der Suchalgorithmus den Zusammenhang der Wörter erkannt hat. Nach diesem Suchvorgang ergab sich beispielsweise bei der IEEE Xplore Digital Library mit 112 Quellen ein überschaubares Ergebnis. Wurde der Suchbegriff allerdings im Abstract verwendet, so ergab dies bereits eine Menge von 876 Quellen. Die erste Filterung der Artikel erfolgte direkt nach der Durchführung der Suche. Hierbei wurde zunächst der Titel gelesen. Weckte der Titel der Publikation das Interesse, so wurde das Abstract gelesen. Allerdings war das Vorgehen nicht Zielführend, da durch die Publikationen keine Einordnung in das Data Mining herausgelesen werden konnte, weswegen zunächst eine Rückwärtssuche anhand der bereits bestehenden Literatur durchgeführt wurde. Hieraus ergaben sich neue Schlagwörter wie „Data preprocessing in Data Mining“, oder „Data transformation in Data Mining“. Diese Schlagwörter wurden in die Datenbanken Springer Link, IEEE Xplore und Scopus eingegeben. Stellte sich eine Publikation anhand von Titel und Abstract als interessant heraus, so wurde zunächst nach dem Schlagwort „transformation“ innerhalb des Textes gesucht. Zudem wurden nun auch Titel mit den Begriffen Normalisierung und Diskretisierung näher betrachtet. Des Weiteren wurde die bereits gefundene Literatur ebenso bearbeitet und anhand des Inhaltsverzeichnisses Werke aussortiert, welche sich nicht mit der Einordnung

der Daten beschäftigen. Nachdem durch eine erste Analyse die genutzten Verfahren herauskristallisierten, so wurde bestehende Literatur hinzugezogen, durch die die einzelnen Verfahren genauer beschrieben werden können. Eine Gesamtübersicht der genutzten Quellen befindet sich unter Anhang A.

## 3.2 Datentransformation zur Knowledge Discovery in Databases

*Tabelle 2 Vergleich der bekanntesten KDD-Modelle*

| Model | Fayyad *et al.* | Cabena *et al.* | Anand & Buchner | CRISP-DM | Cios *et al.* | Generic model |
|---|---|---|---|---|---|---|
| Area | Academic | Industrial | Academic | Industrial | Academic | N/A |
| No of steps | 9 | 5 | 8 | 6 | 6 | 6 |
| Refs | (Fayyad *et al.*, 1996d) | (Cabena *et al.*, 1998) | (Anand & Buchner, 1998) | (Shearer, 2000) | (Cios *et al.*, 2000) | N/A |
| Steps | 1 Developing and Understanding of the Application Domain | 1 Business Objectives Determination | 1 Human Resource Identification | 1 Business Understanding | 1 Understanding the Problem Domain | 1 Application Domain Understanding |
| | | | 2 Problem Specification | | | |
| | 2 Creating a Target Data Set | 2 Data Preparation | 3 Data Prospecting | 2 Data Understanding | 2 Understanding the Data | 2 Data Understanding |
| | | | 4 Domain Knowledge Elicitation | | | |
| | 3 Data Cleaning and Preprocessing | | 5 Methodology Identification | 3 Data Preparation | 3 Preparation of the Data | 3 Data Preparation and Identification of DM Technology |
| | 4 Data Reduction and Projection | | 6 Data Preprocessing | | | |
| | 5 Choosing the DM Task | | | | | |
| | 6 Choosing the DM Algorithm | | | | | |
| | 7 DM | 3 DM | 7 Pattern Discovery | 4 Modeling | 4 DM | 4 DM |
| | 8 Interpreting Mined Patterns | 4 Domain Knowledge Elicitation | 8 Knowledge Post-processing | 5 Evaluation | 5 Evaluation of the Discovered Knowledge | 5 Evaluation |
| | 9 Consolidating Discovered Knowledge | 5 Assimilation of Knowledge | | 6 Deployment | 6 Using the Discovered Knowledge | 6 Knowledge Consolidation and Deployment |

*Tabelle 3 Tabellenauszug der genannten Verfahren zur Datentransformation*

| Autor | Titel | Genannte Vorgehen |
|---|---|---|
| **Baskar et al. (2013)** | A Systematic Approach on Data Preprocessing In Data Mining | Normalisierung, Aggregation |
| **Cleve und Lämmel (2020)** | Data Mining | Normalisierung, Aggregation, Datenglättung, Diskretisierung Transformation zur Dimensionsreduktion, Attributskonstruktion |
| **Alsadi und Bhaya (2017)** | Review od Data Preprocessing Techniques in Data Mining | Datenglättung, Aggregation, Generalisierung, Normalisierung |
| **Garcia et al. (2015)** | Data Preprocessing in Data Mining | Normalisierung, Transformation zur Reduktion |
| **Han et al. (2012)** | Data Mining: Concepts and Techniques | Datenglättung, Attributskonstruktion, Aggregation, Normalisierung, Diskretisierung, Generalisierung |
| **Runkler (2020)** | Data Analysis | Standardisierung |
| **Schulz et al. (2022)** | Analytics in der Industrie | Normalisierung, Aggregation, Diskretisierung |
| **Tamilsevi et al. (2015)** | An efficient preprocessing and postprogressing techniques in Data | Datenglättung, Normalisierung, Aggregation, Generalisierung |
| **Tomar et al. (2014)** | A Survey on Pre-processing and Post-processing Techniques in Data Mining | Datenglättung, Normalisierung, Aggregation |
| **Sangeetha und Sathappan (2018)** | Preprocessing Using Attribute Selection in Data Stream Mining | Datenglättung, Attributskonstruktion, Normalisierung, Diskretisierung, Generalisierung |
| **Larose und Larose (2014)** | Discovering Knowledge in Data | Normalisierung, |
| **Petersohn (2009)** | Data Mining | Aggregation, Normalisierung, Standardisierung und Skalierung |

Die Entdeckung von Wissen in Datenbanken (KDD) bezeichnet den Prozess der Mustersuche in umfangreichen Datenbanken (Klösgen 1996). Dabei werden versteckte Zusammenhänge und Strukturen aufgespürt, die sich in den großen Datenmengen verbergen (Klösgen 1996). Zu diesem Zweck wurden verschiedene Modelle aufgestellt, um Wissen zu generieren. In Kapitel 2.1.2 wurden bereits die zwei bekanntesten Modelle vorgestellt. Kurgan und Musilek (2006) haben die Phasen dieser Modelle gegenübergestellt und mit drei weiteren Modellen verglichen. Fayyad et al. (1996) beschreiben das Vorgehen zur Wissensentdeckung, als iterative und sich wiederholende Prozesse. Bei diesem Modell wird die Datentransformation als

eigenständige Phase gesehen, in welcher der Datensatz speziell auf die genutzte Data Mining-Verfahren anpasst wird (Fayyad et al. 1996). Dies wird in Tabelle 2 unter Schritt 4 dargestellt. Mit der Entwicklung weiterer Modelle wurde diese Phase in die Datenvorverarbeitung (Data Preparation/Data Preprocessing) aufgenommen.

Allerdings sehen alle Modelle die Datenstransformation als den Schritt an, in welchem die Daten für die Data Mining Algorithmen angepasst werden (Kurgan und Musilek 2006). Dieses Phänomen ergibt sich bei der Betrachtung weiterer Publikationen, welche in Tabelle 2 aufgeführt sind. Soweit sich nicht direkt auf das CRISP-DM Modell bezogen wird, stellen die Autoren zunächst die Wissensentdeckung nach Fayyad dar, um im Folgenden die Datentransformation als einen Teil der Vorverarbeitung anzusehen. Daher wird die Datentransformation in dieser Arbeit als ein Schritt innerhalb der Datenvorverarbeitung betrachtet. Innerhalb der Datentransformation, als Anpassung des Datenbestandes an den folgenden Data Mining Algorithmus, zählen die Autoren aus Tabelle 3 eine unterschiedliche Anzahl an Verfahren auf. Es wird allerdings die Herangehensweise bekräftigt, durch Datenransformation die Daten verfahrensadäquat aufzubereiten.

Runkler (2020) beschreibt unter Datentransformation lediglich Standardisierungsverfahren. Diese Ansicht wird von Baskar et al (2013) geteilt, welche als Standardisierung die Normalisierung nennen, um den Datensatz auf einen Wertebereich zwischen 0 und 1 zu skalieren. Die Normalisierung wird von allen Publikationen als gemeinsamer Nenner als Datentransformationsverfahren genannt. Hierbei wird die Standardisierung und Normalisierung wie nach Cleve und Lämmel (2020) als Synonym verwendet.

Tomar et al. (2013) nehmen als weiteres Verfahren die Aggregation hinzu. Diese Ansicht wird von den sechs weiteren Autoren geteilt. Nur Larose und Larose (2014) nennen dieses Verfahren in ihrer Publikation nicht. Baskar et al. (2013) sehen Aggregation zudem als Teil der Datenreduktion, um irrelevante Daten zu entfernen. Schulz et al. (2022) behandeln die Datenaggregation als Ableitung von gegebenen Attributen. Ebenso kann die Attributskonstruktion hierzu gezählt werden (Sangeetha und Sathappan 2018). Dies beruht auf der Beschreibung von Cleve und Lämmel (2020), anhand der gegebenen Attribute neue zu generieren, welches dem Verfahren der Aggregation sehr ähnelt.

Ebenfalls mit 7 Nennungen wir die Datenglättung als weiteres Verfahren zur Datentransformation von den Autoren Tomar et al. (2014), García et al. (2015), Tamilselvi et al. (2015), Alsadi und Bhaya (2017), Han et al. (2012), Sangeetha und Sathappan (2018) sowie Cleve und Lämmel (2020) genannt. Dies hat das bereits genannte Ziel, den Wertebereich durch Clustern, Regression oder Binning zu reduzieren.

Generalisierung wird von fünf Autoren als Datentransformationsverfahren genannt, wobei Han et al. (2012) und Sangeetha und Sathappan (2018) das *Concept hierarchy generation* zur Generalisierung angeben. Alsadi und Bhaya (2017), sowie Tamilselvi (2015) nennen dieses Vorgehen als eine beispielhafte Methode um die Daten zu generalisieren. Han et al. (2012) nennen als Ergebnis der *Concept hierarchy generation* eine Generalisierung der Daten, weswegen es in dieser Arbeit als Methode zur Generalisierung eingeordnet wird.

Als weitere Verfahren der Datenvorverarbeitung nennen die Autoren Diskretierung. Allerdings sind sich die Autoren bei der Zuordnung in die Schritte Datentransformation oder Datenreduktion uneinig. Baskar et al. (2013), Tamilselvi et al. (2015) und Alsadi und Bhaya (2017) zählen die Diskretisierung als Verfahren der Datenreduktion. Dies begründen sie darin, dass die Menge der Daten reduziert wird. Schulz et al. (2022), García et al. (2015), Han et al. (2012), sowie Cleve und Lämmel (2020) zählen die Diskretisierung als Teil der Datentransformation, da es als Verfahren genutzt wird, um metrische Daten in nominale oder ordinale Daten umzuwandeln.

Garcia et al. (2015) zählen allgemein die Verfahren der Transformation zur Dimensionsreduktion zur Datentransformation. Die Autoren Cleve und Lämmel (2020) schneiden die *Principal Component Analysis* (PCA) an, welche von Garcia et al. (2015) ebenso als eine Mehode in

der Datentransformation gesehen wird. Allerdings sehen sie die PCA-Methode auch als Möglichkeit der Datenreduktion. Außerdem wird dieses Verfahren wird von den anderen Autoren, wie Baskar et al. (2013) und Tomar et al. (2014) als Verfahren zur Dimensionsreduktion gesehen und daher der Datenreduktion zugeordnet.

Die Diskretisierung und Principal Component Analysis zeigen den fließenden Übergang zwischen den einzelnen Schritten. Durch einige Verfahren, die das Ziel verfolgen einen einheitlichen Datensatz zu generieren, wird dieser als Nebeneffekt reduziert.

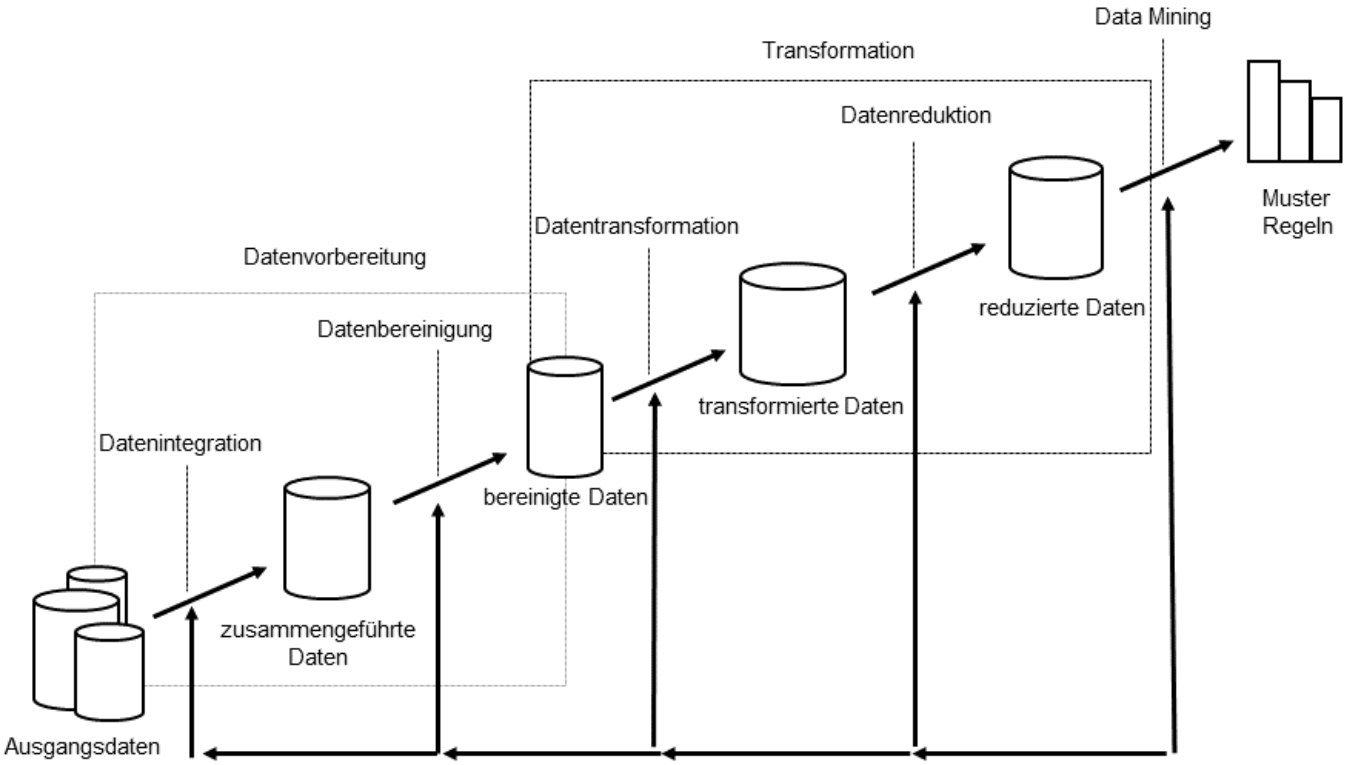


*Abbildung 8 Aufbau Datenvorverarbeitung zur Knowledge Discovery*

Tamilselvi et al. (2015) unterteilt die Datenvorverarbeitung in die Kategorien Vorbereitung und Transformation, welche allgemein die Schritte Datenintegration, Datenbereinigung Datentransformation und Datenreduktion beinhalten. Allerdings ist hervorzuheben, dass Datentransformationen bereits während der Integration und Bereinigung vorgenommen werden können (Cleve und Lämmel 2020). Alt (2020) definiert den Begriff Transformation als einen Prozess, bei dem durch die Umwandlung von Form, Struktur oder Gestalt ein Ausgangs- in einen Zielzustand überführt wird. Dies stützt die Einordnung der Vorbereitungsschritte von Tamilselvi et al. (2015) in die Kategorien Vorverarbeitung und Transformation. Daraus abgeleitet, ordnet diese Arbeit die Schritte Datenintegration und Datenbereinigung zur Datenvorverarbeitung. Der Aufbau im Zusammenhang mit dem darauffolgenden Data Mining wird in Abbildung 7 dargestellt. Die Grundstruktur beruht auf dem Aufbau von Fayyad et al. (1996). Allerdings wurden die Schritte Datenintegration und Datenreduktion hinzugefügt. Anhand der Definition von Alt (2020) werden die Verfahren die Datentransformation, sowie die Datenreduktion in die Kategorie Transformation eingeteilt, da sich in diesen Schritten zum einen die Form durch Erweiterung oder Reduktion von Attributen verändert, zum anderen die Gestalt durch Veränderungen der Werte umgewandelt werden.

Der Datenreduktion werden die Verfahren zugeordnet, welche in Abhängigkeit an das folgende Data Mining Verfahren, zur Reduktion des Datensatzes beitragen und diesen in seiner Form verändern, wobei der Informationsverlust minimiert wird (Han et al. 2012). Dabei werden die Verfahren in die Kategorien der transformationsbasierten und auswahlbasierten Reduktion

eingeteilt (Tomar und Agarwal 2014). Durch Verfahren, wie der Principal Component Analysis werden die Daten ebenso transformiert. Allerdings werden diese Verfahren in der Literatur der Datenreduktion zugeordnet und daher nicht weiter betrachtet.

Petersohn (2009) verfolgt im Gegensatz zu den anderen Autoren den Ansatz die Datenvorverarbeitung in verfahrensabhängig und verfahrensunabhängig einteilt. Unter Verfahrensunabhängigen Datenaufbereitungen, nennt er die Datenanreicherung, Datenreduktion, Stichproben, Aggregation, Dimensionsreduktion und die Behandlung fehlerhafter und fehlender Merkmale (Petersohn 2009). Als Verfahrensabhängige Vorbereitungen nennt Petersohn (2009) unter anderem die Verfahren zur Skalierung, Standardisierung und Normalisierung. Dieser Ansatz der Einordnung wird für die Datentransformation nicht weiterverfolgt.

## 3.3 Datentransformationsverfahren

Zur Datentransformation gehören die Verfahren, welche dazu beitragen, einen Datensatz, um neue Attribute zu erweitern oder gegebene Daten in ihrer Darstellung umzuwandeln, sodass der Datensatz von den folgenden Data Mining ausgewertet werden kann. Jedoch kann dies auch unabhängig des darauffolgenden Verfahrens geschehen. (Cleve und Lämmel 2020). Cleve und Lämmel (2020) unterteilen daher die Datentransformation in verfahrensabhängig und verfahrensunabhängig. Diese Aufteilung wird in Abbildung 9 auf Seite 24 dargestellt. Verfahrensunabhängige Datentransformationen werden genutzt, um einen konsistenten Datenbestand zu erzeugen (Cleve und Lämmel 2020). Hierzu zählen Cleve und Lämmel (2020) die Umwandlungen

- Anpassungen von
    - Datentypen und Konvertierungen
    - Zeichenketten
    - Datumsangaben und Maßeinheiten
- Ableitungen von gegebenen Attributen
    - Kombination oder Separierung von Attributen
    - Berechnung abgeleiteter Werte

Dies überschneidet sich mit der Einteilung von Petersohn (2009). Die Datenglättung und Datenaggregation zählen Cleve und Lämmel (2020) in ihrer Beschreibung als verfahrensabhängige Datentransformationen auf. Allerdings verweisen die Autoren in ihrer Auflistung zu den verfahrensunabhängigen Transformationen ebenso auf diese beiden Verfahren. Bei Aggregationen werden bestehende Werte durch die Berechnung von Summen, Mittelwerten, Medianen oder anderen statistische Verfahren zusammengefasst (Alasadi und Bhaya 2017). Dies bedeutet, dass vorgegebene Daten benötigt werden, aus denen neue Daten abgeleitet werden. Beispielsweise wird die Produktionsanzahl eines Produktes pro Tag auf eine Woche oder Monat hochgerechnet. Aufgrund der Tatsache, dass Aggregationen genutzt werden, um eine neue Aggregationsebene zu erreichen und nicht direkt benötigt wird, um einen bestimmten Mining-Algorithmus anzuwenden wird die Datenaggregation in dieser Arbeit als verfahrensunabhängig angesehen. Da durch die Datenglättung idealisierte Werte erzeugt werden, welche keinen direkten Einfluss auf die genutzten Algorithmen besitzt, wird die Datenglättung aus dem gleichen Grund den verfahrensunabhängigen Datentransformationen zugeordnet. Ebenso wird die Generalisierung genutzt, um, eine höhere Ebene der Granularität zu erreichen (Han et al. 2012). Beispielsweise werden Vertriebspunkte, denen eine Straße zugeordnet wird, zu bestimmten Regionen oder Länder zusammengefasst. Es werden wie bei der Aggregation aus gegebenen Daten, neue Daten abgeleitet. Daher werden Datenaggregation und Generalisierung in Abbildung 9 unter Ableitungen dargestellt.

Neben den verfahrensunabhängigen Datentransformationen, werden die Daten durch verfahrensabhängige Datentransformationen in ein angemessenes Format überführt, mit welchen die Data Mining Verfahren ihre Analysen durchführen können (Cleve und Lämmel 2020). Zu diesen Verfahren zählen Cleve und Lämmel (2020) die Verfahren zur Standardisierung und Normalisierung, sowie die Diskretisierung. Dies überschneidet sich erneut mit der Zuordnung

von Petersohn (2009), welcher Standardisierung und Normalisierung ebenso als verfahrensabhängig betrachtet.

Zusammenfassend lässt sich Datentransformation wie folgt definieren:

Datentransformation ist ein Schritt in der Datenvorverarbeitung zur Knowledge Discoverry in Databases mit dem einen Datensatz in ein für folgende Data Mining Verfahren passende Form überführt wird. Zu diesem Zweck wird der Datensatz durch verfahrensunabhängige Transformationen in einen konsistenten Datensatz überführt, woraufhin verfahrensabhängige Transformationen einen verfahrensadäquaten Datensatz generieren.

In den folgenden Kapiteln werden die Verfahren zur Datenlättung, Standardisierung und Normalisierung, sowie zur Diskretisierung genauer dargestellt. Hierzu werden zunächst die genutzten Quellen dargestellt. Auch wenn eine Quelle ein Verfahren nicht direkt der Datentransformation zugeordnet hat, so wurden die Informationen aus anderen Kapitel genutzt, um die verfahren besser darzustellen.

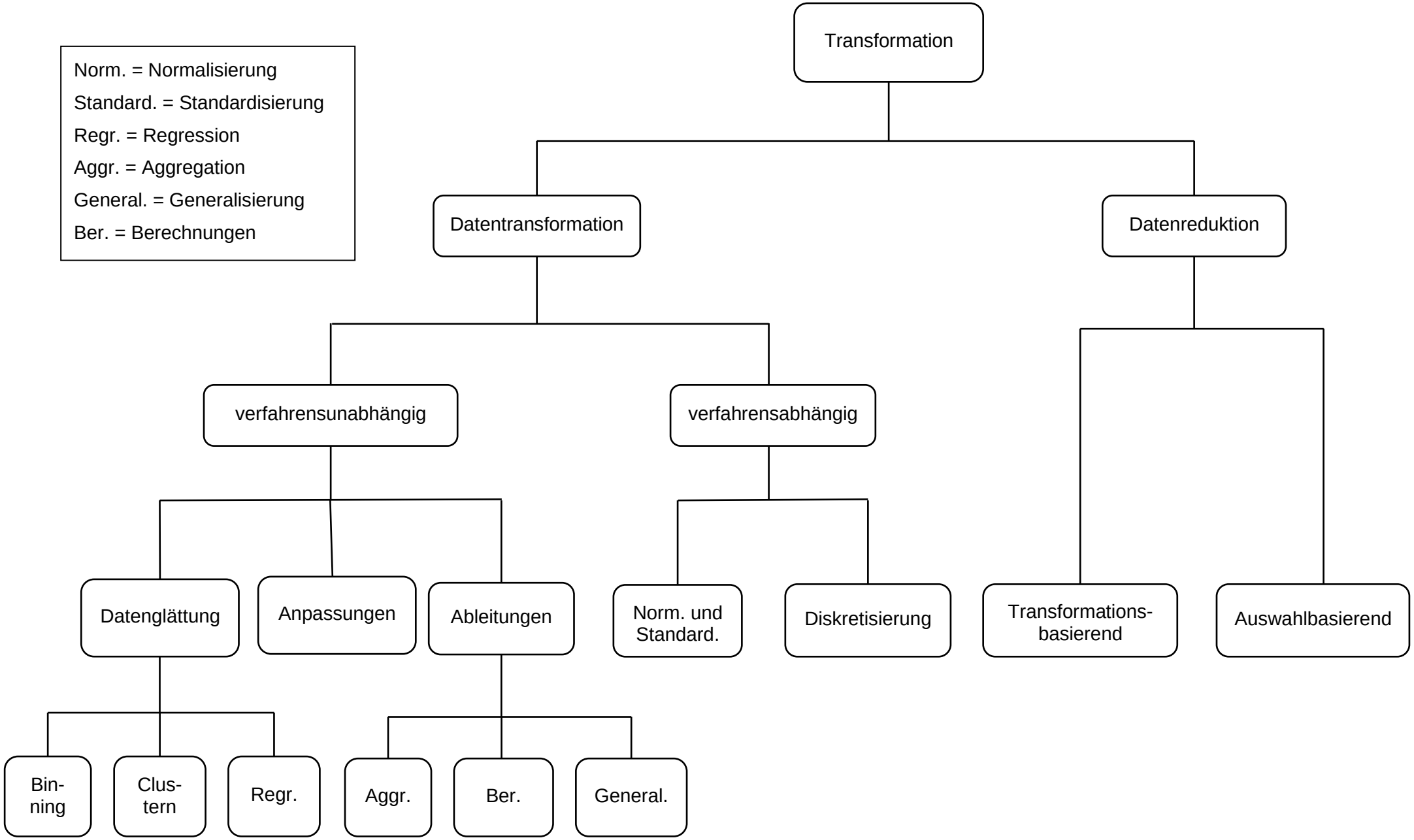


*Abbildung 9 Übersicht der Methoden zur Datentransformation in Bezug zur Kategorie Transformation*

### 3.3.1 Standardisierung und Normalisierung

*Tabelle 4 Quellenübersicht zur Standardisierung und Normalisierung*

| Autor | Titel | genannte Vorgehen |
|---|---|---|
| **Runkler (2020)** | Data Analysis | Standardisierung |
| **Garcia et al. (2015)** | Data Preprocessing in Data Mining | Normalisierung, Transformation zur Reduktion |
| **Schulz et al. (2022)** | Analytics in der Industrie | Normalisierung, Aggregation, Glätttung |
| **Larose und Larose (2014)** | Discovering Knowledge in Data | Normalisierung |
| **Cleve und Lämmel (2020)** | Data Mining | Normalisierung, Aggregation, Datenglättung, Diskretisierung Transformation zur Dimensionsreduktion |
| **Han et al. (2012)** | Data Mining: Concepts and Techniques | Datenglättung, Attributskonstruktion, Aggregation, Normalisierung, Diskretisierung, Generalisierung |
| **Al Shalabi und Shaaban (2006)** | Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix | Normalisierung |
| **Petersohn (2009)** | Data Mining | Aggregation, Normalisierung, Standardisierung und Skalierung |

Verschiedene Attribute können stark unterschiedliche Wertbereiche besitzen (García et al. 2015). Beispielsweise werden bei einem Produktionsprozess verschiedene Produkte hergestellt, deren Lieferzeit eines in Tagen gemessen werden, während die Produktionskosten in Euro angegeben werden. Die Lieferzeit könnte dabei zwischen 1 und 10 Tagen liegen, während die Produktionskosten erheblich zwischen wenigen Hundert bis zu mehreren Tausend Euro variieren. Wenn solche Eigenschaften gemeinsam verwendet werden, können inkorrekte Ergebnisse erzielt werden, da die Bereiche der Eigenschaften so unterschiedlich sind (Runkler 2020). In der linken Ansicht von Abbildung 10 zeigt Runkler (2020) einen zufällig generierten Datensatz, indem eine zweidimensionale Gaußsche Verteilung mit dem Mittelwert μ = (30000, 100) und der Standardabweichung s = (9000, 30) verwendet wurde. Der Datensatz erscheint als eine horizontale Linie nahe bei Null. Nach Runkler (2020) kann dies eine valide Darstellung der Werte sein. Allerdings ist es in Fällen, bei denen die Attribute als gleich wichtig betrachtet werden, nützlich, die Daten auf einen ähnlichen Bereich zu transformieren (Runkler 2020). Um solche Darstellungen zu erreichen, werden die Verfahren Normalisierung oder Standardisierung genutzt (Schulz et al. 2022).

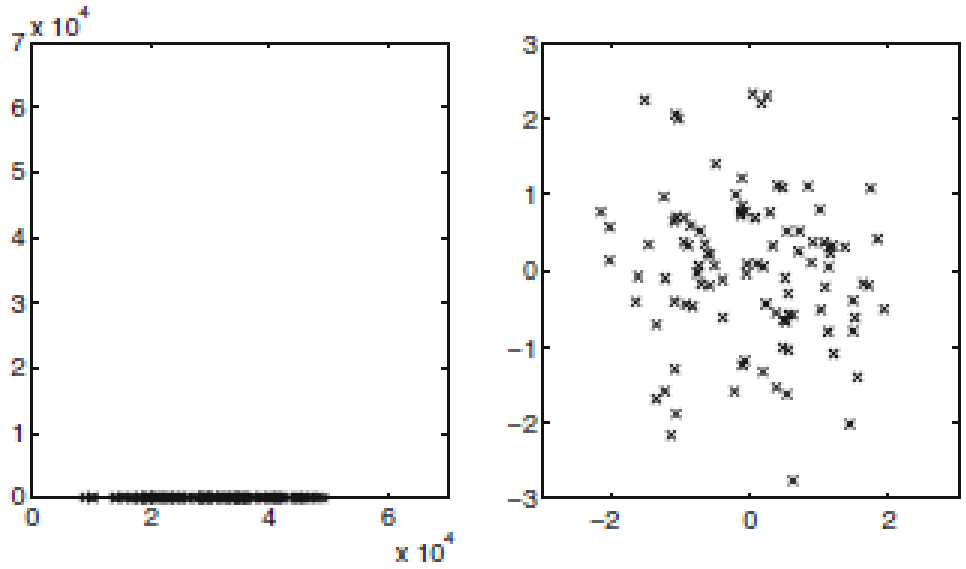


*Abbildung 10 Normalisierung eines Datensatzes*

Normalisierung oder Normierung ist wie bereits beschrieben eine Art der Skalierung, bei der die Werte auf ein bestimmtes Intervall skaliert werden. Im Kontext der Knowledge Discovery nennen die Autoren einen Wertebereich zwischen [-1,1] oder [0,1] (Han et al. 2012). Zur Skalierung auf ein bestimmtes Intervall wird Min-Max Normalisierung, Z-Wert-Normalisierung oder Dezimalskalierung genutzt (Larose und Larose 2014). Cleve und Lämmel (2020) nennen zudem die Logarithmische Skalierung, bei der alle Zahlen in einen zugehörigen Logarithmus mit der Basis B umgewandelt werden.

**Min-Max-Normalisierung** führt eine lineare Transformation der Originaldaten durch (Al Shalabi und Shaaban 2006). Durch Min-Max-Normalisierung werden Abstände zwischen den ursprünglichen Datenwerten eingehalten (Han et al. 2012). Zur Min-Max-Normalisierung auf das Intervall [0,1] werden von einem Attribut A lediglich das Maximum max($A_i$) und das Minimum min($A_i$) benötigt (Cleve und Lämmel 2020). Der normierte Wert berechnet sich nach Larose und Larose (2014) wie folgt

$$X_{neu} = \frac{X - \min(A_i)}{\max(A_i) - \min(A_i)}$$

Zuerst wird von allen Werten der Minimalwert subtrahiert, wodurch das Minimum auf 0 gesetzt wird (Cleve und Lämmel 2020). Danach erfolgt eine Division durch die maximale Differenz zwischen zwei Werten (Cleve und Lämmel 2020).

Wenn die Werte auf ein beliebiges Intervall [a,b] skaliert werden sollen, so nutzen Cleve und Lämmel (2020) folgende Formel

$$X_{neu} = (b - a)\frac{X - \min(A_i)}{\max(A_i) - \min(A_i)} - a$$

In manchen Fällen ist die Min-Max-Normalisierung nicht durchführbar, wenn beispielsweise das Minimum und Maximum nicht bekannt sind (García et al. 2015). Auch wenn Minimum und Maximum bekannt sind, können Ausreißer die die Min-Max-Normalisierung beeinflussen (García et al. 2015).

Durch **Z- Wert -Normalisierung** werden die Daten basierend auf Mittelwert und Standardabweichung des Attributes A normalisiert (Al Shalabi und Shaaban 2006). Durch diese Transformation werden die Werte mit einem Mittelwert von 0 und einer Standardabweichung von 1 dargestellt (García et al. 2015). Petersohn (2009) verwendet folgende Formel

$$X_{neu} = \frac{X - \bar{A}}{\sigma_A} -$$

Die Z-Wert-Normalisierung berechnet für ein Attribut A den statistischen Mittelwert $\bar{A}$ und die Standardabweichung $\sigma_A$ der Attributwerte (Cleve und Lämmel 2020). Anschließend wird von jedem Wert der Mittelwert subtrahiert, und das Ergebnis wird durch die Standardabweichung dividiert (Larose und Larose 2014). Durch die Skalierung auf einen Mittelwert 0 und einer Standardabweichung von 0 wird dieses Verfahren auch Z-Wert-Standardisierung genannt und daher der Standardisierung zugeordnet (Petersohn 2009).

Eine Abwandlung der Z-Wert-Normalisierung ersetzt die Standardabweichung durch die mittlere absolute Abweichung (Han et al. 2012). Diese berechnet sich nach García et al. (2015) nach der Formel

$$s_A = \frac{1}{n}\sum_{i=1}^{n} |X_i - \bar{A}|$$

Als Ergebnis der mittleren absoluten Abweichung wird nach Han (2012) die Z-Wert-Normalisierung mit der folgenden Formel durchgeführt

$$X_{neu} = \frac{X - \bar{A}}{s_A} -$$

Vorteil dieser Berechnung liegt darin, dass Ausreißer einen geringeren Einfluss auf das Ergebnis haben, da die Abstände vom Mittelwert als Betrag genommen und nicht wie bei der Standardabweichung quadriert werden (García et al. 2015).

Ein einfacher Weg, numerische Werte zu normalisieren ist die **Dezimalskalierung**, bei der die Dezimalkommastelle per Division durch 10 verschoben wird (García et al. 2015).

Die Anzahl der zu verschiebenden Kommastellen hängt vom absoluten Maximum ab (Al Shalabi und Shaaban 2006). Larose und Larose (2014) beschreiben die Berechnung mit folgender Formel

$$X_{neu} = \frac{X}{10^j}$$

Wobei j der kleinste Wert ist, sodass max($|X_{neu}|$) < 1 (Han et al. 2012).

Dieses Verfahren stellt sicher, dass die normalisierten Werte zwischen -1 und 1 liegen (Larose und Larose 2014).

Han et al. (2012) merken an, dass die Originaldaten durch Normalisierung ein bisschen verändert werden können. Daher ist es wichtig, die genutzten Parameter abzuspeichern, um zukünftige Daten in der gleichen Weise zu bearbeiten (Al Shalabi und Shaaban 2006).

### 3.3.2 Datenglättung

*Tabelle 5 Quellenübersicht zur Datenglättung*

| Autor | Titel | Verwendete Verfahren |
|---|---|---|
| **Cleve und Lämmel (2020)** | Data Mining Knowledge Discovery in Databases | Normalisierung, Aggregation, Datenglättung, Diskretisierung Transformation zur Dimensionsreduktion |
| **Han et al. (2012)** | Data Mining: Concepts and Techniques | Datenglättung, Attributskonstruktion, Aggregation, Normalisierung, Diskretisierung, Generalisierung |
| **Larose und Larose (2014)** | Discovering Knowledge in Data | Normalisierung, Datenglättung |
| **Alsadi und Bhaya (2017)** | Review od Data Preprocessing Techniques in Data Mining | Datenglättung, Aggregation, Generalisierung, Normalisierung |
| **Bramer et al. (2020)** | Principles of Data Mining | Clustern |

Datenglättung soll die Wertemenge des Datensatzes reduzieren, um so bessere Lösungen zu erhalten (Cleve und Lämmel 2020). Durch die Methoden Eingruppierung (Binning), Clustering und Regression sollen die Werte aus der Datenmenge durch idealisierte Werte ersetzt werden (Cleve und Lämmel 2020). Diese wurde grob im Kapitel 2.1.3 beschrieben.

**Clusteranalyse**

Clusteranalyse oder einfach Clustering bezeichnet den Prozess, eine Gruppe von Datenobjekten in Untermengen aufzuteilen (Han et al. 2012). Die daraus entstehenden *Cluster* sind Ansammlungen mit sich ähnelnden Werte, welche sich von anderen Werten in anderen Clustern unterscheiden (Larose und Larose 2014).

Cleve und Lämmel (2020) und Han et al. (2012) überschneiden sich in der Einteilung der Klassen in:

- Partitionierende Clusterbildung
- Hierarchische Clusterbildung
- Dichtebasierte Clusterbildung

Eine **Partitionierungsmethode** nimmt eine Menge von n Objekten und erstellt k Aufteilungen der Daten, wobei jede Aufteilung einen Cluster repräsentiert und $k \leq n$ ist (Han et al. 2012). Dabei werden die Daten in k Gruppen aufgeteilt, wobei jede Gruppe mindestens ein Objekt enthalten muss (Han et al. 2012). Bei dieser Methode werden eine beliebige Anzahl an Aufteilungen gewählt (Cleve und Lämmel 2020). Die Objekte in den Aufteilungen werden daraufhin zwischen den Clustern umgeordnet bis alle Objekte den Clustern zugeordnet sind, denen sie am ähnlichsten sind (Cleve und Lämmel 2020). Dadurch gehört jedes Objekt genau einer Gruppe an (Han et al. 2012).

Bei der **hierarchischen Clusteranbildung** entsteht durch wiederholte Aufteilung (divisive Methoden) oder Vereinigung (agglomerative Methoden) eine baumartige Clusterstruktur (Dendrogramm) von existierenden Clustern (Larose und Larose 2014). Der agglomerative Ansatz, auch als Bottom-up-Ansatz bekannt, beginnt damit, dass jedes Objekt eine eigenständige Gruppe bildet (Han et al. 2012). Schrittweise werden Objekte oder Gruppen, die nahe beieinander liegen, miteinander verbunden, bis alle Gruppen zu einer einzigen Gruppe verbunden sind oder eine Abbruchbedingung erfüllt ist (Larose und Larose 2014). Bei der wiederholten

Aufteilung, auch als Top-Down-Ansatz bezeichnet, sind Objekte im gleichen Cluster (Han et al. 2012). In jeder aufeinanderfolgenden Iteration wird ein Cluster in kleinere Cluster aufgeteilt, bis schließlich jedes Objekt in einem Cluster oder eine Abbruchbedingung erfüllt ist (Larose und Larose 2014).

Die **dichtebasierenden Clusteranalysen** erstellen Cluster, bei denen in der Umgebung eines Objekts eine Mindestanzahl an weiteren Objekten vorhanden ist (Cleve und Lämmel 2020). Beispielsweise sollte für jeden Datenpunkt innerhalb eines definierten Clusters die Umgebung eines bestimmten Radius mindestens eine Mindestanzahl von Punkten enthalten (Han et al. 2012).

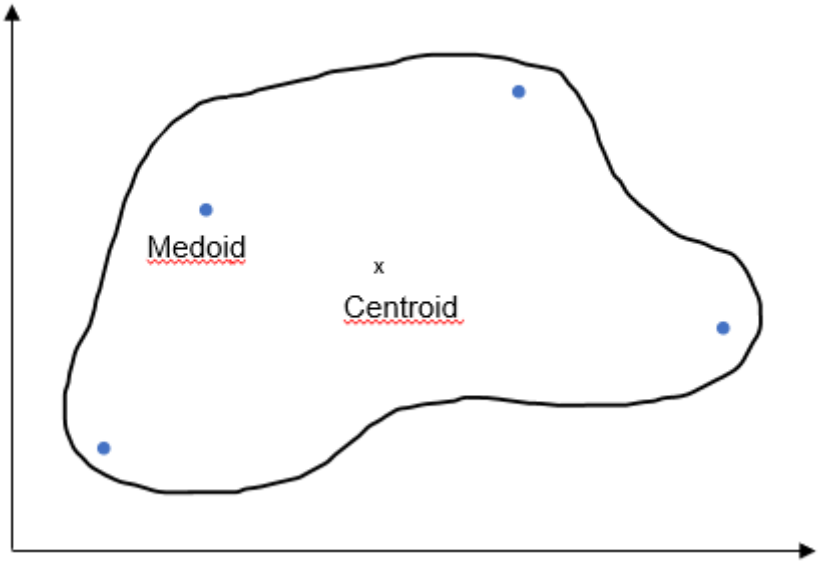


*Abbildung 11 Beispielhafte Clusterbildung*

Ein Beispiel für die Clusteranalyse, wie in Abbildung 11 dargestellt, ist der k-Means-Algorithmus. Bei dieser Methode wird zu Beginn bestimmt, wie viele Cluster gebildet werden sollen (Bramer 2020). Anhand dieser Anzahl k werden k Werte als zentrale Punkte der Cluster (Centroiden) ausgewählt (Larose und Larose 2014). Die Datenpunkte, welche den zentralen Punkten der Cluster am nächsten liegen, werden den jeweiligen Clustern zugeordnet (Bramer 2020). Diese werden als Medoiden bezeichnet (Cleve und Lämmel 2020). In den entstandenen Clustern werden daraufhin neue Centroiden berechnet (Cleve und Lämmel 2020). Dies geschieht beispielsweise durch die Berechnung der Mittelwert der X- und Y-Werte (Cleve und Lämmel 2020). Daraufhin wird jeder Wert der neue Centroid und dem jeweiligen Cluster zugeordnet und der nächste Centroid berechnet (Larose und Larose 2014). Dieses Vorgehen wiederholt sich solange, bis sich die Werte der Centroiden nicht weiter verändern (Bramer 2020).

**Binning**

Durch Binning werden sortierte Daten durch Einbezug der umliegenden Daten geglättet (Han et al. 2012), Die Daten werden dabei sogenannten Bins zugeordnet (Alasadi und Bhaya 2017). Bei Glätten nach Mittelwert-Bin werden die Werte in den Bins durch den Mittelwert ersetzt (Alasadi und Bhaya 2017). Beim Glätten durch Median Bin werden die Daten durch den Median ersetzt (Han et al. 2012). Beim Glätten der Daten durch Grenzwert-Bin werden die Daten innerhalb eines Bins mit den Grenzwerten verglichen und durch jenen ersetzt, welcher näher zum Ausganswert liegt (Alasadi und Bhaya 2017).

Als Beispiel soll der Datensatz $X = \{31,36,41,44,48,,49,51,55,57,58,61,67,71,73,78\}$ durch drei Bins geglättet werden. Dazu werden die Werte in drei Bins eingeteilt.

- Bin 1: 31,36,41,44,48
- Bin 2: 49,51,55,57,58
- Bin 3: 61,67,71,73,78

Glätten durch Mittelwert-Bin ergibt das Ergebnis

- Bin 1: 50,50,50,50
- Bin 2: 54,54,54,54
- Bin 3: 70,70,70,70

Glätten durch Median-Bin ergibt das Ergebnis

- Bin 1: 41,41,41,41,41
- Bin 2: 55,55,55,55,55
- Bin 3: 71,71,71,71,71

Glätten durch Grenzwert-Bin ergibt das Ergebnis

- Bin 1: 31,31,48,48,48
- Bin 2: 49,49,58,58,58
- Bin 3: 61,61,78,78,78

Das Vorgehen zur Einteilung wird im folgenden Kapitel näher erläutert. Allgemein wird der Glättungseffekt größer je breiter die Bins gestaltet sind (Alasadi und Bhaya 2017). Binning wird auch zur Diskretisierung genutzt. Dieses Verfahren wird im folgenden Kapitel behandelt.

### 3.3.3 Diskretisierung

*Tabelle 6 Quellenübersicht zur Diskretisierung*

| Autor | Titel | Verfahren |
|---|---|---|
| **Cleve und Lämmel (2020)** | Data Mining | Normalisierung, Aggregation, Datenglättung, Diskretisierung Transformation zur Dimensionsreduktion |
| **Bakar et al. (2009)** | Building A New Taxonomy For Data Discretization Techniques | Diskretisierung |
| **Garcia et al. (2013)** | A Survey of Discretization Techniques Taxonomy and Empirical Analysis_in Supervised Learning | Diskretisierung |
| **Baskar et al. (2013)** | A Systematic Approach on Data Pre-processing In Data Mining | Diskretisierung als Reduktion |
| **Aggarwal (2015)** | Discovering Knowledge in Data | Normalisierung, Standardisierung, Diskretisierung |
| **Han et al. (2012)** | Data Mining: Concepts and Techniques | Datenglättung, Attributskonstruktion, Aggregation, Normalisierung, Diskretisierung, Generalisierung |
| **Larose und Larose (2014)** | Discovering Knowledge in Data | Binning als Analysemethode |

Viele Verfahren zum Data Mining können nummerische Werte nicht auswerten und benötigen eine kategorische Einteilung (Cleve und Lämmel 2020). Daher werden bei der Diskretisierung die numerischen Ausprägungen in Intervalle zusammengefasst (Baskar et al. 2013). Ordinale Diskretisierungen wandeln metrische Daten in ordinale qualitative Daten, wohingegen die nominale Diskretisierung die Daten in nominale Daten umwandeln und das Ordnungsmaß nicht berücksichtigen (Garcia et al. 2013).

Bis April 2013 wurden bereits mindesten 80 Diskretisierungsmethoden veröffentlicht (Garcia et al. 2013). Daher werden die Diskretisierungsverfahren in Abbildung 11 auf Seite 33 nach Bakar et al. (2009) kategorisiert. Die Auswahl der Diskretisierungsmethode hängt zum einen vom Datensatz, als auch vom Hintergrundwissen des Analysten ab (Bakar et al. 2009). Beim nicht-hierarchischen Ansatz legt der Experte eine angemessene Anzahl an Bins fest (Bakar et al. 2009). Ist das Wissen nicht vorhanden werden hierarchische Ansätze genutzt, bei denen Aufteilungs- Zusammenführungs- oder eine Kombinationstechnik angewendet werden, um die Zuverlässigkeit der Intervalle zu gewährleisten (Bakar et al. 2009). Aufteilungsmethoden (Splits) setzen einen Schnittpunkt zwischen allen möglichen Grenzpunkten und teilen den Bereich in zwei Intervalle auf (Garcia et al. 2013). Dieser Ansatz wird auch Top-down-Diskretisierung genannt (Baskar et al. 2013). Im Gegensatz dazu beginnen Zusammenführungsmethoden (Merge), auch Bottom-up-Diskretisierung genannt, mit einer vordefinierten Aufteilung und entfernen einen Schnittpunkt, um beide benachbarten Intervalle zu vermischen (Garcia et al. 2013). Bei der Kombinationstechnik werden Splits und Merges abwechselnd angewendet (Garcia et al. 2013). Außerdem gibt es Diskretisierungsmethoden, welche mehrere Intervalle gleichzeitig splitten oder morgen können (Garcia et al. 2013).

Die zweite Ebene teilt die Diskretisierungen in überwachte (SV) und unüberwachte (USV) Methoden ein. Überwachte Diskretisierungen können genutzt werden, sobald eine die gegebenen Daten in Klassen eingeteilt werden können (Cleve und Lämmel 2020). Gibt es diese Einteilung nicht, so müssen unüberwachte Verfahren eingesetzt werden (Baskar et al. 2013).

Zur Diskretisierung eignen sich Verfahren, wie beispielsweise Clusteranalysen oder Binningmethoden (Cleve und Lämmel 2020). Die verbreitetste Methode ist jedoch das Binning (Han et al. 2012). Bei diesem Verfahren werden im Gegensatz zur Datenglättung die Werte einer bestimmten Anzahl von Bins zugeordnet und durch die neuen Kategorien ersetzt (Han et al. 2012). Binningmethoden sind beispielsweise das e*qual width binning* und das e*qual frequenzy binning* (Larose und Larose 2014).

Beim **equal width binning** wird eine Anzahl an Kategorien durch den Analysten vorgegeben (Larose und Larose 2014). Der Bereich [a,b] wird so festgelegt, dass die Differenz b − a für alle Bereiche gleich ist (Aggarwal 2015). Dieser Ansatz hat den Nachteil, dass er nicht für Datensätze funktioniert, die ungleichmäßig über die verschiedenen Bereiche verteilt sind (Aggarwal 2015).

Beim **equal frequenzy binning** werden der vorgegebenen Anzahl an Kategorien k gleich viele Einträge mit der Anzahl k/n zugewiesen, wobei n die gesamt Anzahl an Werten ist (Larose und Larose 2014). Das Attribut wird sortiert und in Gleichgroße Bereiche unterteilt (Aggarwal 2015). Die Teilungspunkte werden auf Basis der sortierten Attributswerte ausgewählt, so dass jeder Bereich dieselbe Anzahl von Datensätzen enthält (Aggarwal 2015).

Als Beispiel soll der Datensatz $X = \{1,1,1,2,2,3,3,12,12,13,13,44\}$ in die drei Kategorien Niedrig, Mittel und Hoch eingeteilt werden.

Nach dem **equal width binning** wird der Wertebereich in fünfzehnerschritte eingeteilt und die Werte wie folgt zugeordnet

- Niedrig: $0 \leq X < 15$ enthält die Werte $X = \{1,1,1,2,2,3,3,12,12,13,13\}$
- Mittel: $15 \leq X < 30$ enthält keine Werte
- Hoch: $30 \leq X < 45$ enthält die Werte $X = \{44\}$

Wie an diesem Beispiel erkennbar, ist diese Methode stark von Ausreißern beeinflussbar. Auch hat die Vorgabe der Binanzahl einen hohen Einfluss auf das Ergebnis (Han et al. 2012).

Nach **equal frequenzy binning** ist $n = 12, k = 3$ und $n/k = 4$. Also werden jedem Bin vier Werte zugeordnet.

- Niedrig enthält vier Werte $X = \{1,1,1,2\}$
- Mittel enthält vier Werte $X = \{2,3,3,12\}$
- Hoch enhält vier Werte $X = \{12,13,13,44\}$

Hier ist zu beachten, dass ein Wert aus dem Bin N*iedrig*, ebenfalls dem Bin *Mittel* ein Wert aus dem Bin M*ittel* ebenfalls dem Bin H*och* zugeordnet ist. Allerdings sollten gleiche Werte den gleichen Kategorien zugeordnet werden (Larose und Larose 2014).

Durch Diskretisierung geht der Detaiillierungsgrad verloren, da die Daten generalisiert werden (Baskar et al. 2013). Allerdings fügen Baskar et al. (2013) hinzu, dass die Daten durch dieses Verfahren aussagekräftiger und dadurch einfacher zu interpretieren sind.

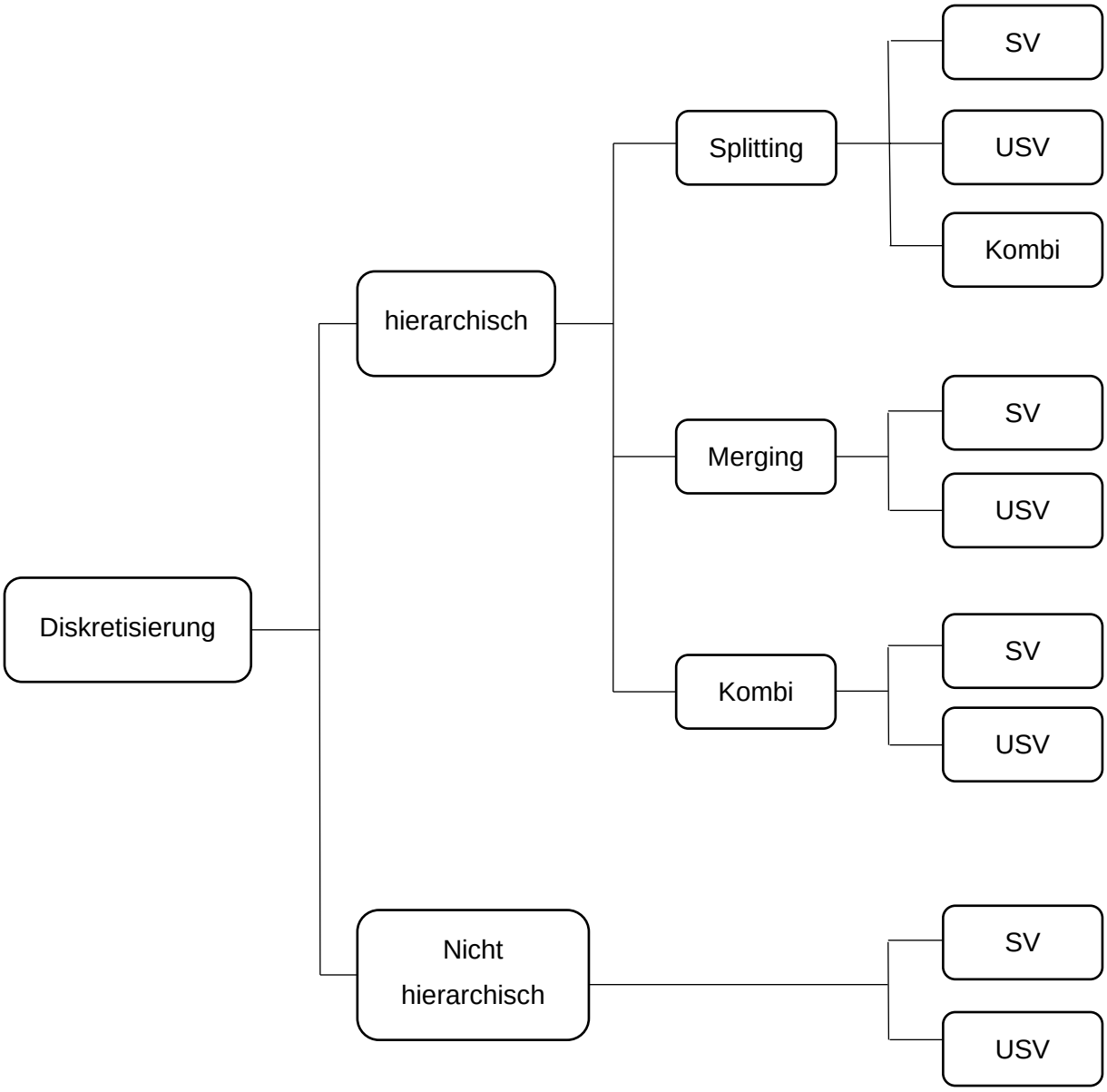


*Abbildung 12 Kategorien zu Diskretisierung*

## 3.4 Diskussion und Fazit

Zum Abschluss dieser Arbeit, werden die Ergebnisse zusammengefasst und in Bezug auf die Forschungsfrage interpretiert und diskutiert. Zu diesem Zweck setzt sich dieser Teil kritisch mit der vorrangegangenen Arbeit auseinander und zeigt die positiven als auch negativen Aspekte auf. Außerdem befasst sich dieses Kapitel mit den Beschränkungen, welche diese Arbeit umfassen und wie sie in die Literatur eingeordnet werden kann.

In dieser Arbeit wurde durch eine systematische Literaturrecherche die Datentransformation im Kontext der Datenvorverarbeitung zur Knowledge Discovery in Databases ermittelt. Anhand der Ergebnisse konnte dies Datentransformation allgemein eingeordnet werden und durch einen exemplarischen Vergleich grundlegende Verfahren herausgearbeitet werden, welche in der Literatur der Datentransformation zugeschrieben werden. Durch diese Ergebnisse konnte ein allgemeines Verständnis durch eine abschließende Definition erlangt werden.

Zu Beginn der Ausarbeitung wurde im Kapitel 2 der Stand der Technik dargestellt, um ein allgemeines Verständnis der Knowledge Discovery, unter verstärkter Berücksichtigung der Da-

tentransformation zu erhalten und allgemeingültige Verfahren darzustellen, die unter Datentransformation zu finden sind. Anhand dieser Darstellung wurden Suchbegriffe abgeleitet, mit denen eine systematische Literaturrecherche durch verschiedene Kombinationen der Begriffe, sowie Zuordnungen in Suchfeldern durchgeführt. Allerdings ergab diese Form der Suche vorerst keine zielführenden Ergebnisse, weswegen anhand der vorhandenen Literatur durch Rückwärtssuche neue Quellen und ein zweiter Suchansatz gefunden werden konnten. Allerdings ist auch bei dem zweiten Suchvorgang die Anzahl der gefundenen Quellen zu kritisieren. Da sich allerdings der gefundene Rahmen mit den bereits vorhandenen Quellen deckte, wurde mit der Auswertung begonnen. In dieser ergab sich, dass sich die Autoren zunächst auf das Vorgehensmodell von Fayyad berufen, allerdings die Phase Datentransformation im Verlauf der Darstellung als Teil der Datenvorverarbeitung nennen und die Schritte Datentransformation und Datenreduktion aufteilen. Durch den exemplarischen Vergleich wurden dazu die Verfahren zur Datenglättung, Aggregation, Generalisierung, Standardisierung und Normalisierung, sowie die Diskretisierung zugeordnet. Verglichen mit dem Stand der der Technik, ist hervorzuheben, dass Verfahren, welche die Datentypen anpassen, oder Attribute im Datensatz kombinieren oder separieren von den Autoren nicht berücksichtigt werden. Da es sich allerdings auch um Transformationen der Daten handelt, wurden diese in dieser Arbeit mit zu den Datentransformationen gezählt, unter Berücksichtigung, dass diese Verfahren bereits während der Datenintegration durchgeführt werden können. Des Weiteren behandelt außer den Autoren Cleve und Lämmel kein Autor die Umwandlung von ordinalen oder nominalen Daten in metrische Daten. Es ist zudem hervorzuheben, dass je nach Veröffentlichung verschiedene Verfahren genannt wurden. Allerdings wurden die Verfahren Aggregation, sowie Standardisierung und Normalisierung, von den meisten Autoren genannt. Zu Aggregationen gehören statistische Methoden, wie dem Berechnen des Mittelwertes oder Medians. Standardisierung und Normalisierung werden häufig als Synonyme genutzt und dienen zur Anpassung der Werte in einen bestimmten Wertebereich. Hier werden die Verfahren Min-Max-Normalisierung, Z-Wert-Normalisierung oder Dezimalskalierung eingesetzt. Datenglättungsverfahren, welche zur Datentransformation genutzt werden, haben im Gegensatz zu der Bereinigung, in welcher Rauschen oder Ausreißer geglättet werden, das Ziel, idealisierte Werte zu erhalten. Zu diesem Verfahren gehören Techniken, welche dem Clustern, der Regression oder dem Binning zugeordnet werden. Die Regression wird in die Kategorien lineare und multidimensionale Regression eingeteilt. Das Clustern und Binning findet sich zudem in den Verfahren zur Diskretisierung wieder, wobei die Literatur das Binning hervorhebt. Clusteranalysen lassen sich in die Kategorien partitionierende, hierarchische und dichtebasierende Methoden einteilen. Beim Binning können Methoden wie das equal width oder equal frequenzy Binning genutzt werden, um Daten in sogenannte Buckets oder Bins einzuteilen. Dazu werden für die Datenglättung die Werte in den erzeugten Bins durch den Mittelwert, Median, oder den näheren Grenzwert ersetzt. Es stellte sich beim Vergleich der Zuordnungen der Verfahren zu Datentransformation oder Datenreduktion heraus, dass je nach Ziel und Interpretation der Verfahren, diese sowohl der Datentransformation als auch der Datenreduktion zugeordnet werden können. So begründen beispielsweise Baskar et al., sowie Bhaya et al., dass durch Diskretisierung die Anzahl an Werten reduziert wird, weswegen sie das Verfahren zur Datenreduktion zählen. Cleve und Lämmel sehen in diesem Verfahren das Ziel metrische Daten in Kategorien einzuteilen, um so Data Mining Verfahren anwenden zu können, welche lediglich ordinale oder nominale Daten auswerten. Allgemein kann gesagt werden, dass bereits durch Verfahren wie der Aggregation, der Generalisierung oder der schon genannten Diskretisierung der Datensatz verkleinert wird wird und so eine Datenreduktion stattfindet. Wie in Kapitel 3.2 dargestellt wurde, werden ebenfalls Transformationen der Daten zur Datenreduktion genutzt, welche unter der Kategorie transformationsbasierte Reduktionen eingeteilt werden. Zudem wird durch Filtern der Attribute die Form des Datensatzes verändert, welches der allgemeinen Definition von Transformation entspricht. Allerdings hat dies das klare Ziel die Dimension zu reduzieren, weswegen diese Transformationen nicht zur Datentransformationen zählen. Durch den Vergleich der Ergebnisse in Kapitel 3 und der Datenvorverarbeitung aus Kapitel 2.1.3, ist zu erkennen, dass Verfahrensübergreifend in der Datentransformation, als auch in der gesamten

Datenvorverarbeitung die gleichen Methoden angewendet werden. Daher kann von einem fließenden Übergang zwischen den Schritten gesprochen werden. Anhand dieser Überschneidungen kann die Datenvorverarbeitung in die Kategorien Datenvorverarbeitung und Transformation eingeteilt werden. welche sich wiederrum in Datenintegration, Datenbereinigung, Datentransformation und Datenreduktion aufteilen. Daher konnte in dieser Arbeit eine weitere Form der Kategorisierung der Schritte zur Datenvorverarbeitung ermittelt werden. In dieser Aufteilung ist das Ziel der Datentransformation den Datensatz in ein für die Data Mining Phase verfahrensadäquates Format anzupassen. Nach Cleve und Lämmel kann dies geschehen, um einen konsistenten Datensatz zu erhalten, oder einen Datensatz zu generieren, welcher ein Format aufweist, mit welchen die Algorithmen der Data Mining Verfahren rechnen können. Werden die einzelnen Verfahren genauer betrachtet, so ist erkennbar, dass die Verfahren der Datenglättung, Aggregation, Generalisierung und Typanpassung nicht genutzt werden, um den von Cleve und Lämmel genannten konsistenten Datensatz zu erhalten, weswegen sie den verfahrensunabhängigen Transformationen zugeordnet werden. Standardisierung und Normierung, sowie die Diskretisierung werden allerdings dann angewendet, wenn der Datensatz speziell angepasst werden muss. Dadurch werden diese Verfahren den verfahrensabhängigen Transformationen zugeordnet. Durch diese Einteilung konnte anhand dieser Ausarbeitung eine weitere Kategorisierung der Datentransformationen vorgenommen werden. Da bei der Generalisierung, sowie der Aggregation vorhandene Werte genutzt werden, um neue Werte abzuleiten, wurden diese Verfahren der Kategorie Ableitungen zugeordnet, zu welcher auch die direkte Berechnung neuer Werte zweier Attribute gezählt wird. So konnten in dieser Arbeit drei Kategorisierungsansätze herausgearbeitet werden, in welche sich die Datentransformation zuordnen und einteilen lässt. Zudem konnte anhand der Ergebnisse eine detaillierte Definition der Datentransformation formuliert werden.

Wird die durchgeführte Literaturrecherche und die daraus resultierenden Ergebnisse kritisch betrachtet, so können mehrere Limitationen dieser Bachelorarbeit hervorgehoben werden. Besonders in Bezug auf die Domäne ist hervorzuheben, dass hier keine Praxisbezogenen Quellen zur Ausarbeitung hinzugezogen wurden. Zudem ergab sich aus den Quellen keine Tendenz der genutzten Data Mining Verfahren in Produktion und Logistik, um die Verfahren der Datentransformationen einzuschränken. Allerdings ist ebenso zu betonen, dass Produktion und Logistik ein sehr großes Feld ist und die in Unternehmen durchgeführten Analysen stark variieren können. Hierzu ist es von Vorteil eine Übersicht zu besitzen, in welcher die wichtigsten Verfahren aufgeführt sind, um ein Verständnis der Datentransformation zu erhalten und darauf aufbauend, das Wissen zu vertiefen. In Bezug auf die durchgeführte Literaturrecherche ist die Anzahl der genutzten Quellen zu kritisieren. Innerhalb der Recherche wurden weitere Verfahren zur Transformation, wie die Transformation zur Normalität von Larose und Larose, gefunden. Allerdings war die Informationsdichte zu diesem Verfahren nicht gegeben. Mit einer weiteren Untersuchung kann hier Klarheit geschaffen werden. Trotz der geringen Menge an Quellen, konnten jedoch eine Schnittmenge von Verfahren zur Datentransformation ermittelt und dargestellt werden. Daher können aus dieser Arbeit neue Erkenntnisse gewonnen werden. Es wurde gezeigt, dass die Transformation den Datensatz an die folgenden Data Mining-Verfahren anpasst. Allerdings wird innerhalb der Datentransformation zwischen verfahrensabhängig und verfahrensunabhängig unterschieden. So kann in Unternehmen zunächst anhand der Ziele die Anforderungen an einen Datensatz abgeleitet und dieser darauf angepasst werden, um im darauffolgenden Schritt den Datensatz an die jeweiligen Data Mining-Verfahren anzupassen. Dies ist sinnvoll, da nach dem CRISP-DM-Modell in der Modellierungsphase mehrere Modelle genutzt werden, welche unterschiedliche Anforderungen besitzen können.

Insgesamt lässt sich sagen, dass durch diese Arbeit ein grundlegendes Verständnis der Datentransformation im Kontext der Datenvorverarbeitung erlangt werden konnte. Dies spiegelt sich in der ausformulierten Definition von Datentransformation wider.

# 4 Zusammenfassung und Ausblick

Das Ziel dieser Arbeit war es, ein Verständnis des Begriffes der Datentransformation im Kontext der Datenvorverarbeitung in Produktion und Logistik zu erlangen. Um dieses Ziel zu erreichen, wurde in Kapitel 2 der Stand der Technik dargestellt. Dieser bestand aus der Einführung in die Wissensentdeckung in Datenbanken, in welcher die zwei bekanntesten Modelle vorgestellt wurden, um daraufhin die Phase der Datenvorverarbeitung genauer darzustellen. Außerdem wurden allgemeine Verfahren vorgestellt, welche zur Datentransformation genutzt werden. Im dritten Kapitel wurde die systematische Untersuchung der Datentransformation umgesetzt. Hierzu wurde zu Beginn das Vorgehen zur systematischen Literaturrecherche nach vom Brocke et al. (2009) erläutert, um dieses Vorgehen mit der Taxonomie nach Cooper (1988) anzuwenden. Zu diesem Zweck wurden aus dem zweiten Kapitel Schlüsselbegriffe abgeleitet und der Rechercheumfang festgelegt. Das Ergebnis dieser Literaturrecherche wurde in einer Gesamtübersicht der genutzten Quellen unter Nennung der genutzten Verfahren erstellt. Anhand dieser Übersicht wurden die Quellen miteinander verglichen. Hierzu wurde zunächst die Einordnung der Datentransformation in der Knowledge Discovery betrachtet, Darauf aufbauend wurden die Quellen auf Gemeinsamkeiten und Unterschiede in den genutzten Verfahren überprüft. Hierzu wurden die genutzten Quellen aus der Gesamtübersicht exemplarisch dargestellt. Auf diesen Erkenntnissen aufbauend, wurde die Datentransformation innerhalb der Datenvorverarbeitung betrachtet und mit den dazugehörigen Schritten der Datenvorverarbeitung eingeordnet. Dabei konnte festgestellt werden, dass die Verfahren ebenfalls in anderen Schritten angewendet werden können und je nach Interpretation zugeordnet werden. Anhand der Gemeinsamkeiten des Vergleiches wurden die Verfahren definiert, welche der Datentransformation zugeordnet werden. Hierzu wurden diese Verfahren in weitere Kategorien unterteilt und in einer Graphik visuell dargestellt, Um das Hauptziel zu erreichen wurde im Abschluss des Kapitels 3.3 eine Definition der Datentransformation formuliert. Im Anschluss daran wurden in den Kapiteln 3.3.1 bis 3.3.3 die Verfahren genauer dargestellt. Zu diesem Zweck wurden zu Beginn der Kapitel die Auszüge der Gesamtübersicht tabellarisch dargestellt und verschiedene Methoden vorgestellt, welche für die Verfahren angewendet werden können.

Es ist in dieser Arbeit zu erkennen, dass die Darstellung der Datentransformation im Kontext der Datenvorverarbeitung verbessert werden kann. Anhand der Literaturrecherche wurden keine direkten Anforderungen des Data Minings an die Datentransformation innerhalb der Domäne Produktion und Logistik abgeleitet. Daher wird vorgeschlagen, anhand von weiteren Untersuchungen wie Umfragen in der Industrie oder weiteren systematischen Literaturrecherchen angewendete Data Mining Verfahren in Produktion und Logistik herauszuarbeiten, um darauf aufbauend Anforderungen abzuleiten, welche Datentransformationen in dieser Domäne angewendet werden müssen. Zudem können weitere Recherchen zu Methoden der einzelnen Verfahren anhand eines anderen Suchansatzes in der Datentransformation durchgeführt werden. Anhand der gewonnenen Erkenntnisse kann ein Entscheidungsmodell erstellt werden, welche durch Input der gegeben und benötigten Daten passende Methoden verschlägt. Zudem gibt es viele weitere Transformationsansätze, welche in dieser Arbeit nicht behandelt wurden. Daher ist zu empfehlen, in diesem Gebiet weiter zu forschen.

## Literaturverzeichnis


Aggarwal, Charu C. (2015): Data Mining. The Textbook. 1st ed. 2015. Cham: Springer International Publishing; Imprint: Springer.

Aggarwal, Charu C. (2017): Outlier Analysis. 2nd ed. 2017. Cham: Springer International Publishing; Imprint: Springer.

Al Shalabi, Luai; Shaaban, Zyad (2006): Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix, S. 207–214. DOI: 10.1109/DEPCOS-RELCOMEX.2006.38.

Alasadi, Suad A.; Bhaya, Wesam S. (2017): Review of Data Preprocessing Techniques in Data Mining.

Bakar, Azuraliza Abu; Othman, Zulaiha Ali; Shuib, Nor Liyana Mohd (2009): Building a new taxonomy for data discretization techniques, S. 132–140. DOI: 10.1109/DMO.2009.5341896.

Baskar, S. S.; Arockiam, L.; Charles, S. (2013): A Systematic Approach on Data Pre-processing In Data Mining. In: *Compusoft*, S. 335–339.

Baumann, S.; Gnisia, M.; Feifel, P.; Klingauf, U. (2018): Identifikation und Behandlung von Ausreißern in Flugbetriebsdaten für Machine Learning Modelle. Online verfügbar unter https://www.dglr.de/publikationen/2018/480169.pdf, zuletzt geprüft am 02.11.2023.

Bensberg, Frank (2001): Web Log Mining als Instrument der Marketingforschung. Ein systemgestaltender Ansatz für internetbasierte Märkte. Gabler Edition Wissenschaft. Wiesbaden: Deutscher Universitätsverlag (Informationsmanagement und Controlling).

Bodendorf, Freimut (2006): Daten- und Wissensmanagement. 2., aktualisierte und erweiterte Auflage. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg (Springer-Lehrbuch).

Bramer, Max (2020): Principles of Data Mining. London: Springer London.

Brauckmann, Otto (2019): Digitale Revolution in der industriellen Fertigung – Denkansätze. 1. Aufl. 2019. Berlin, Heidelberg: Springer Berlin Heidelberg. Online verfügbar unter http://nbn-resolving.org/urn:nbn:de:bsz:31-epflicht-1582081.

Cleve, Jürgen; Lämmel, Uwe (2020): Data Mining. 3rd. Boston: De Gruyter.

Cooper, Harris M. (1988): Organizing Knowledge Syntheses: A Taxonomie of Literature Reviews. In: *Knowledge in Society*, S. 104–126.

Dommermuth, Maximilian (2020): Entwicklung und Anwendung eines konsekutiven integralen Transformationskonzeptes für Werke von Industrieunternehmen mit variantenreicher Fertigung. Dissertation.

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.

Garcia, Salvador; Luengo, J.; Sáez, José Antonio; López, Victoria; Herrera, F. (2013): A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. In: *IEEE Trans. Knowl. Data Eng.* 25 (4), S. 734–750. DOI: 10.1109/TKDE.2012.35.

García, Salvador; Herrera, Francisco; Luengo, Julián (2015): Data Preprocessing in Data Mining. 1st ed. 2015. Cham: Springer International Publishing; Imprint: Springer (Intelligent Systems Reference Library, 72).

Grün, Oskar; Jammernegg, Werner (2019): Grundzüge der Beschaffung, Produktion und Logistik. 4., aktualisierte Auflage. Hg. v. Sebastian Kummer. Hallbergmoos: Pearson (wi. Wirtschaft).

Günther, Hans-Otto; Tempelmeier, Horst (2013): Produktion und Logistik. Supply Chain und Operations Management. 10., erw. und verb. Aufl. Norderstedt: Books on Demand.

Han, Jiawei; Pei, Jian; Kamber, Micheline (2012): Data mining. Concepts and techniques. 3rd ed. Amsterdam, Boston, Heidelberg, Amsterdam, Boston, Heidelberg: Morgan Kaufmann; Elsevier (The Morgan Kaufmann series in data management systems).

Huber, Andreas; Laverentz, Klaus (2019): Logistik. 2., überarbeitete und korrigierte Auflage. München: Verlag Franz Vahlen (Vahlens Kurzlehrbücher). Online verfügbar unter https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1933263.

Klösgen, Willi (1996): Knowledge Discovery in Databases and Data Mining. In: *Foundations of Intelligent Systems*, S. 623–632. Online verfügbar unter https://link.springer.com/book/10.1007/3-540-61286-6?page=4#toc, zuletzt geprüft am 06.12.2023.

Küpper, Hans-Ulrich; Helber, Stefan (2004): Ablauforganisation in Produktion und Logistik. 3., überab. und erw. Aufl. Stuttgart: Schäffer-Poeschel.

Kurgan, Lukasz; Musilek, Petr (2006): A survey of Knowledge Discovery and Data Mining process models. In: *Knowledge Eng. Review* 21, S. 1–24. DOI: 10.1017/S0269888906000737.

Larose, Daniel T.; Larose, Chantal D. (2014): Discovering knowledge in data. An introduction to data mining. Second edition. Hoboken: Wiley (Wiley series on methods and applications in data mining). Online verfügbar unter https://ieeexplore.ieee.org/servlet/opac?bknumber=10066951.

Luengo, Julián; García-Gil, Diego; Ramírez-Gallego, Sergio; García, Salvador; Herrera, Francisco (2020): Big Data Preprocessing. Enabling Smart Data. 1st ed. 2020. Cham: Springer International Publishing; Imprint Springer (Springer eBook Collection).

Martin, Heinrich (2016): Transport- und Lagerlogistik. Systematik, Planung, Einsatz und Wirtschaftlichkeit. 10. Auflage. Wiesbaden: Springer Vieweg.

Mertens, Peter; Bodendorf, Freimut; König, Wolfgang; Picot, Arnold; Schuhmann, Matthias; Hess, Thomas (2012): Grundzüge der Wirtschaftsinformatik. 11. Aufl. Berlin, Heidelberg: Springer Gabler (Springer-Lehrbuch).

Mishra, Puneet; Biancolillo, Alessandra; Roger, Jean Michel; Marini, Federico; Rutledge, Douglas N. (2020): New data preprocessing trends based on ensemble of multiple preprocessing techniques. In: *TrAC Trends in Analytical Chemistry* 132, S. 116045. DOI: 10.1016/j.trac.2020.116045.

Müller, Roland M.; Lenz, Hans-Joachim (2013): Business Intelligence. Berlin, Heidelberg: Springer (eXamen.press).

North, Klaus (2011): Wissensorientierte Unternehmensführung. Wertschöpfung durch Wissen. 5., aktualisierte und erw. Aufl. Wiesbaden: Gabler Verlag / Springer Fachmedien Wiesbaden GmbH Wiesbaden (Gabler Lehrbuch).

Petersohn, Helge (2009): Data Mining. Verfahren, Prozesse, Anwendungsarchitektur. Zugl.: Leipzig, Univ., Habil, 2004. München, Wien: Oldenbourg. Online verfügbar unter http://www.degruyter.com/search?f_0=isbnissn&q_0=9783486593334&searchTitles=true.

Pfohl, Hans-Christian (2018): Logistiksysteme. Betriebswirtschaftliche Grundlagen. 9. Aufl. 2018. Berlin, Heidelberg: Springer Berlin Heidelberg. Online verfügbar unter http://nbn-resolving.org/urn:nbn:de:bsz:31-epflicht-1611506.

Runkler, Thomas A. (2020): Data Analytics. Models and Algorithms for Intelligent Data Analysis. Third Edition. Wiesbaden: Springer Fachmedien Wiesbaden; Springer Vieweg (Springer eBook Collection).

Sangeetha, R.; Sathappan, S. (2018): Proceedings of the 3rd International Conference on Communication and Electronics Systems (ICCES 2018). 15-16, October 2018. Piscataway, NJ: IEEE. Online verfügbar unter https://ieeexplore.ieee.org/servlet/opac?punumber=8717974.

Schuh, Günther; Zeller, Violett; Stich, Volker (Hg.) (2022): Digitalisierungs- und Informationsmanagement. Springer-Verlag GmbH. Berlin, Heidelberg: Springer Vieweg (Handbuch Produktion und Management, 9). Online verfügbar unter https://swbplus.bsz-bw.de/bsz1759487813cov.htm.

Schulz, Thomas; Ayaz, Boris; Kröckel, Johannes; Huber, Marco; Ingold, Remo; Oppermann, Henrik et al. (2022): Analytics in der Industrie. Schlüsseltechnologie für die digitale Transformation. 1. Auflage. Hg. v. Thomas Schulz. Würzburg: Vogel Communications Group GmbH & Co. KG.

Sharafi, Armin (2013): Knowledge Discovery in Databases. Eine Analyse des Änderungsmanagements in der Produktentwicklung. Zugl.: München, Techn. Univ., Diss., 2012. Wiesbaden: Springer Gabler (Springer Gabler Research).

Sharma, Sumana; Osei-Bryson, Kweku-Muata (2008): Organization-Ontology Based Framework for Implementing the Business Understanding Phase of Data Mining Projects. In: *Proceedings of the Annual Hawaii International Conference on System Sciences*. DOI: 10.1109/HICSS.2008.339.

Tamilselvi, R.; Sivasakthi, B.; Kavitha, R. (2015): An efficient preprocessing and postprocessing techniques in Data Mining. In: *International*, S. 80–85.

Tomar, Divya; Agarwal, Sonali (2014): A Survey on Pre-processing and Post-processing Techniques in Data Mining. In: *IJDTA* 7 (4), S. 99–128. DOI: 10.14257/ijdta.2014.7.4.09.

vom Brocke, Jan; Riemer, Kai; Niehaves, Björn; Plattfaut, Raif (2009): Reconstructing the Giant - On the Importance of Rigour in Documenting the Literature Search Process. In: *In Proceedings of the 17th European Conference on*, S. 2206–2217.

Wirth, R.; Hipp, Jochen (2000): CRISP-DM: Towards a standard process model for data mining. In: *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, S. 29–39. Online verfügbar unter https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf, zuletzt geprüft am 01.11.2023.

# Anhang

## Anhang A: Genutzte Quellen zum Verständnis der Datentransformation

| Autor | Titel | Genannte Vorgehen |
|---|---|---|
| **Baskar et al. (2013)** | A Systematic Approach on Data Preprocessing In Data Mining | Normalisierung, Aggregation |
| **Cleve und Lämmel (2020)** | Data Mining | Normalisierung, Aggregation, Datenglättung, Diskretisierung Transformation zur Dimensionsreduktion, Attributskonstruktion |
| **Alsadi und Bhaya (2017)** | Review od Data Preprocessing Techniques in Data Mining | Datenglättung, Aggregation, Generalisierung, Normalisierung |
| **Garcia et al. (2015)** | Data Preprocessing in Data Mining | Normalisierung, Transformation zur Reduktion |
| **Han et al. (2012)** | Data Mining: Concepts and Techniques | Datenglättung, Attributskonstruktion, Aggregation, Normalisierung, Diskretisierung, Generalisierung |
| **Runkler (2020)** | Data Analysis | Standardisierung |
| **Schulz et al. (2022)** | Analytics in der Industrie | Normalisierung, Aggregation, Diskretisierung |
| **Tamilsevi et al. (2015)** | An efficient preprocessing and postprogressing techniques in Data | Datenglättung, Normalisierung, Aggregation, Generalisierung |
| **Tomar et al. (2014)** | A Survey on Pre-processing and Post-processing Techniques in Data Mining | Datenglättung, Normalisierung, Aggregation |
| **Sangeetha und Sathappan (2018)** | Preprocessing Using Attribute Selection in Data Stream Mining | Datenglättung, Attributskonstruktion, Normalisierung, Diskretisierung, Generalisierung |
| **Petersohn (2009)** | Data Mining | Aggregation, Normalisierung, Standardisierung und Skalierung |
| **Larose und Larose (2014)** | Discovering Knowledge in Data | Normalisierung, Binning und Clustern als Analysemethode |
| **Al Shalabi und Shaaban (2006)** | Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix | Normalisierung |

| Autor | Titel | Genannte Vorgehen |
|---|---|---|
| **Bakar et al. (2009)** | Building A New Taxonomy For Data Discretization Techniques | Diskretisierung |
| **Klösgen (1996)** | Knowledge Discovery in Databases | Einordnung Datentransformation in KDD |
| **Bramer et al. (2020)** | Principles of Data Mining | Clustern als Analyse |
| **Garcia et al. (2013)** | A Survey of Discretization Techniques Taxonomy and Empirical Analysis_in Supervised Learning | Diskretisierung |