technische universität dortmund

IT in Produktion und Logistik

Literaturbasierte Entwicklung einer Entscheidungsunterstützung zur Auswahl von Datentransformationsverfahren als Vorverarbeitung für Data Mining in der Automobilproduktion

Masterarbeit zur Erlangung des Grades M. Sc.

Vorgelegt von:	Abdulsamed Acar
Matrikelnummer:	202074
Studiengang:	Master Maschinenbau

Ausgabedatum:	19.08.2024
Abgabedatum:	14.02.2024

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

Technische Universität Dortmund

Fakultät Maschinenbau

Fachgebiet IT in Produktion und Logistik

Inhaltsverzeichnis

Abbildungsverzeichnis I
Tabellenverzeichnis II
Formelverzeichnis III
Abkürzungsverzeichnis IV
1 Einleitung 1
2 Grundlagen der Datentransformation in der Automobilproduktion 3
2.1 Daten 3
2.1.1 Datenebenen 6
2.1.2 Datenqualität 8
2.2 Data Mining 10
2.3 Methoden des Data Mining 12
2.4 Datentransformation in der Datenvorverarbeitung 17
2.4.1 Datentransformationsverfahren 18
2.4.2 Entscheidungsunterstützung 23
2.5 Automobilproduktion 25
2.5.1 Allgemeines zur Automobilproduktion 25
2.5.2 Daten in der Automobilproduktion 26
2.5.3 Datentransformationsverfahren in der Automobilproduktion 28
3 Systematische Literaturrecherche zu Datentransformationsverfahren 31
3.1 Methodik der systematischen Literaturrecherche 31
3.2 Bedingungen der Literaturauswahl 32
3.3 Durchführung der Literaturrecherche 36
3.4 Kategorisierung der Datentransformationsverfahren 41
4 Entwicklung einer Entscheidungsunterstützung zur Auswahl eines Datentransformationsverfahrens 45
4.1 Kriterien zur Anwendung der Datentransformationsverfahren 45
4.2 Anforderungen des Data Mining an Datentransformationsverfahren 49
4.3 Randbedingungen der Automobilproduktion für Datentransformationsverfahren 51
4.4 Konzipierung der Entscheidungsunterstützung 53
5 Exemplarische Anwendung der Entscheidungsunterstützung 58

5.1 Einführung eines Fallbeispiels 58
5.2 Anwendung der Entscheidungsunterstützung 59
5.3 Ergebnisse 63
6 Diskussion und Fazit 72
7 Zusammenfassung und Ausblick 76
8 Literaturverzeichnis 77
Anhang A: Systematische Literaturrecherche 87
A.1 Übersicht der systematischen Literaturrecherche 87
A.2 Prüfung der Literatur auf die Verfahren 89
A.3 Unbekannte Verfahren aus den Quellen 90
A.4 Rückwärtssuche zur Datentransformation aus den Publikationen der Literaturrecherche 91
Anhang B: Algorithmen der exemplarischen Validierung aus Python 93
B.1 Dicretization der Qualität 93
B.2 Discretization der Prozessparameter und der Qualität 94
B.3 Assoziationsanalyse der Prozessparameter und der Qualität 96
B.4 Clustering ohne Normalization 98
B.5 Clustering mit Normalization 100
B.6 Python Bibliothek 102
Eidesstattliche Versicherung 103

Abbildungsverzeichnis

Abbildung 1: Big-Data-Werkzeuge (in Anlehnung an Badshah et al. (2024), S. 290) 7

Abbildung 2: Initiale Cluster und Bilden der Centroiden (in Anlehnung an Cleve und Lämmel (2024), S. 158) 15

Abbildung 3: Neu berechnetete Centroiden (in Anlehnung an Cleve und Lämmel (2024), S. 158).....16

Abbildung 4: Statistik über die Nutzung der Datentransformationsverfahren pro Quelle 40

Abbildung 5: Entscheidungsunterstützung zur Auswahl von Datentransformationsverfahren als Vorverarbeitung für Data Mining in der Automobilproduktion.....55

Abbildung 6: Pfad der Entscheidungsunterstützung für das erste Szenario.....59

Abbildung 7: Pfad der Entscheidungsunterstützung für das zweite Szenario.....62

Abbildung 8: Datenpunkte der Prozessparameter mit farblicher Visualisierung der Qualität .68

Abbildung 9: Clustering der Datenpunkte ohne Normalization.....69

Abbildung 10: Clustering der Prozessparameter mit Normalization 69

Abbildung 11: Datenpunkte in normalisierten Wertebereichen der Prozessparameter mit farblicher Visualisierung der Qualität 70

Tabellenverzeichnis

Tabelle 1: Standard Datentypen und Java Basisdatentypen (in Anlehnung an Güting und Dieker (2018), S. 42) 4

Tabelle 2: Datenqualitätskategorien und -dimensionen (in Anlehnung an Strong et al. (1997), S. 104) 8

Tabelle 3: Data-Mining-Prozesse (in Anlehnung an Runkler (2015), S. 3) 10

Tabelle 4: Taxonomie (in Anlehnung an Cooper (1988), S. 3-9) 31

Tabelle 5: Kategorisierung der Datentransformationsverfahren 42

Tabelle 6: Kriterien zur Anwendung der Datentransformationsverfahren 47

Tabelle 7: Gegenüberstellung der Kriterien zur Anwendung, Anforderungen des Data Mining und der Randbedingungen der Automobilproduktion 54

Tabelle 8: Externe Transformationen aus der Kategorisierung der Datentransformationsverfahren 60

Tabelle 9: Ausschnitt aus den Ergebnissen der Discretization.....65

Tabelle 10: Ergebnisse der Frequent Itemsets mit einem Support von 30% 65

Tabelle 11: Assoziationsregeln mit Konfidenzwerten 66

Formelverzeichnis

Formel 1: Gini-Index 12
Formel 2: Entropie 13
Formel 3: Informationsgewinn an den Teilknoten 13
Formel 4: Abhängigkeit der Variablen in der Regression 14
Formel 5: Quadratische Regressionsfehler Minimierung 14
Formel 6: Euklidische Distanzberechnung 16
Formel 7: Z-Wert Normalisierung 22
Formel 8: Min-Max-Normalisierung 23
Formel 9: Umwandlung der Normalisierung 23

Abkürzungsverzeichnis

CFK	Carbonfaser-verstärkter Kunststoff
CRISP-DM	Cross-Industry-Standard-Process-for-Data-Mining
EUS	Entscheidungsunterstützungssystem
IoT	Internet-of-Things
KDD	Knowledge-Discovery-in-Database
KI	Künstlicher Intelligenz
OHE	One-Hot-Encoding
SEMMA	Sample, Explore, Modify, Model and Assess
SQL	Structured-Query-Lanuage

1 Einleitung

Laut Kahraman und Haktanır (2023) ist die Digitalisierung in allen Schritten der Automobilindustrie beteiligt und liefert durch die Verbindungen der Produktion Daten. Die Unternehmen streben danach das vollständige Potenzial aus den Daten auszuschöpfen, um Prozesse effektiver zu gestalten (Loebbecke und Picot 2015). Diesbezüglich ist das Data Mining ein wesentlicher Prozess, bei dem riesige Datenmengen extrahiert werden, um durch erkannte Muster die Produktionsprozesse zu optimieren (García et al. 2015).

Damit die Daten sinnvoll genutzt werden können, muss davor eine Datenvorverarbeitung erfolgen (El Morr et al. 2022). Die Datenvorverarbeitung beinhaltet z. B. eine Datenbereinigung, Anpassung und Formatierung, um die generelle Datenqualität zu verbessern (El Morr et al. 2022). Ein wichtiger Schritt der Datenvorverarbeitung ist die Datentransformation. Der analyserelevante Zieldatenbestand wird mithilfe der Datentransformation in das benötigte Schema umgewandelt (Cleve und Lämmel 2024). Die Rohdaten sind nicht alle in dem gewollten Format, sodass sich die Wertebereiche unterscheiden (Runkler 2015). Runkler (2015) behauptet, dass der Anwender der Datentransformationsverfahren den beobachteten und den gewünschten Wertebereich betrachten sollte. Basierend darauf muss die Auswahl des geeigneten Datentransformationsverfahren erfolgen (Runkler 2015). Die Entscheidung hängt nicht nur vom Wertebereich, sondern von mehreren Faktoren ab und kann je nach Anwender variieren. Laut Alpar et al. (2019) sind Systeme der Entscheidungsunterstützung von Bedeutung, da der Mensch entlastet wird und somit eine Rationalitätssteigerung des Entscheidungsprozesses besteht. Alpar et al. (2019) behaupten, dass kognitive Fähigkeiten des Menschen einen Einfluss auf die Entscheidung haben. Die Autoren führen weiter aus, dass die Entscheidungsunterstützung entweder durch subjektive Informationen eines Anwenders oder durch objektive Einflüsse des Standes der Technik entstehen kann.

Das Ziel dieser Arbeit ist die Entwicklung einer Entscheidungsunterstützung zur Auswahl der passenden Datentransformationsverfahren in der Automobilproduktion. Dafür werden mit einer systematischen Literaturrecherche Datentransformationsverfahren festgestellt und die Bedingungen der Nutzung dieser Verfahren thematisiert. Das Hauptziel ist die Entwicklung einer Entscheidungsunterstützung basierend auf einer systematischen Literaturrecherche, um die Auswahl des Datentransformationsverfahren für die Daten der Automobilproduktion zu vereinfachen. Zur Erreichung des Hauptziels werden Teilziele definiert. Das erste Teilziel umfasst die Vorstellung bestimmter Datentransformationsverfahren, um die Grundlage für die systematische Literaturrecherche zu schaffen. Das zweite Teilziel thematisiert die Kategorisierung der Datentransformationsverfahren, die aus der systematischen Literaturrecherche hervorgegangen sind. Mithilfe der systematischen Literaturrecherche und dem Wissen aus dem Stand der Technik werden die Datentransformationsverfahren gruppiert, um einen generellen Überblick über die verschiedenen Verfahren zu erstellen. Das dritte Teilziel ist das Bestimmen der Kriterien zur Nutzung der Datentransformationsverfahren, der Vorgaben des Data Mining und

der Randbedingungen der Automobilproduktion. Im vierten Teilziel wird durch die exemplarische Anwendung die Wirksamkeit der Entscheidungsunterstützung geprüft und ausgewertet.

Um diese Ziele zu erreichen, wird in Kapitel 2 grundlegendes bzgl. der Daten, der Datenvorverarbeitung und des Data Mining vorgestellt. Dabei werden Skalenniveaus und Datenqualität erklärt, die im Hinblick der Datenvorverarbeitung thematisiert werden. Innerhalb dieses Kapitels wird die Verknüpfung zur Datentransformation erstellt, um die ersten Bezugspunkte zu schaffen. Am Ende des Kapitels wird die Automobilindustrie als Domäne eingeleitet und in den Kontext mit den Datentransformationsverfahren gesetzt, um den Anwendungsbereich in der Arbeit zu definieren. Das Kapitel 3 thematisiert die systematische Literaturrecherche. Zuerst wird die allgemeine Methodik beschrieben und im Anschluss mithilfe der erarbeiteten Begriffe aus Kapitel 2 durchgeführt, damit verschiedene Datentransformationsverfahren ermittelt werden. Dabei werden die Begriffe der Datentransformationsverfahren ggf. mit der Domäne zusammengesetzt und mit verschiedenen Suchfiltern benutzt, um nutzbare Ergebnisse zu erzielen. Danach sollen diese Datentransformationsverfahren kategorisiert werden, um im nächsten Kapitel mit der Entwicklung der Entscheidungsunterstützung zu starten. Mithilfe der verschiedenen Kategorien können mehrere Datentransformationsverfahren als Gruppe betrachtet und für diese Kriterien zur Anwendung bestimmt werden. Dazu werden die Anforderungen aus dem Data Mining und die Randbedingungen der Automobilproduktion betrachtet, um abgängig von allen Faktoren die Entscheidungsunterstützung zu konzipieren. In Kapitel 5 wird mithilfe eines Fallbeispiels die konzipierte Entscheidungsunterstützung validiert, um zu prüfen, ob die Entscheidungsunterstützung für den Kontext der Datentransformation in der Automobilproduktion anwendbar ist. Nach der Auswertung der Ergebnisse werden im vorletzten Kapitel mögliche Verbesserungen bzw. Mängel thematisiert und der Entstehungsprozess der Entscheidungsunterstützung diskutiert. Im Fazit wird die finale Entscheidungsunterstützung vorgestellt und für die Anwendung in der Produktion erklärt. In Kapitel 7 werden die Erkenntnisse der Arbeit zusammengefasst und zukünftige Aspekte der Entscheidungsunterstützung von Datentransformationsverfahren thematisiert.

2 Grundlagen der Datentransformation in der Automobilproduktion

In diesem Kapitel werden Daten thematisiert, die im Hinblick auf die Digitalisierung in der Automobilproduktion essenziell sind und die Basis für die Datentransformationsverfahren bilden. Dabei werden Aspekte wie Datenqualität und -ebenen erklärt. In Abschnitt 2.2 folgt die Einführung in das Data Mining. Das Data Mining wird aufgrund der Datenvorverarbeitung thematisiert. Es bestehen Anforderungen im Data Mining, die im vorherigen Schritt, also in der Datenvorverarbeitung erfüllt werden müssen. In Abschnitt 2.3 werden die verschiedenen Methoden des Data Mining vorgestellt. Danach wird in Abschnitt 2.4 die Datentransformation eingeleitet und als Teilschritt der Datenvorverarbeitung dargestellt. Im Unterabschnitt 2.4.1 werden die verschiedenen Datentransformationsverfahren erklärt, die im Fokus dieser Arbeit stehen. Danach werden grundlegende Informationen bzgl. Entscheidungsunterstützungen thematisiert, da diese im späteren Verlauf für die Datentransformationsverfahren von Bedeutung sind. Am Ende in Abschnitt 2.5 wird der Bezug zwischen Datentransformationsverfahren und der Automobilindustrie erstellt. Dabei werden Prozesse und Daten in der Automobilproduktion vorgestellt. Im Unterabschnitt 2.5.3 werden die wesentlichen Bestandteile dieser Arbeit, also die Datentransformation, Automobilproduktion und Entscheidungsunterstützung zusammengeführt.

2.1 Daten

Als Daten werden Zeichen (Buchstaben, Ziffern, Sonderzeichen) definiert, wenn sie eine Ordnung bzw. eine Syntax zugeordnet bekommen (North 2021). North (2021) führt weiter aus, dass Daten noch nicht interpretierte Symbole und somit eine beliebige Aneinanderreihung von Zeichen sind. Vieweg et al. (2012) ergänzen zu der Definition, dass Daten auch Bildpunkte oder Töne sein können. Aufgrund dieser Tatsache können Daten in verschiedenen Formaten vorliegen und müssen durch eine saubere, zentrale Dokumentation verwaltet werden, um einen Mehrwert zu generieren (Hildebrand et al. 2021). Man unterscheidet zwischen unstrukturierten, semistrukturierten und strukturierten Daten (Cleve und Lämmel 2024). Unstrukturierte Daten sind Bilder und Texte (Cleve und Lämmel 2024). Semistrukturierte Daten sind z. B. Informationen auf Web-Seiten, da die Daten in verschiedenen Formaten organisiert sein können (Cleve und Lämmel 2024). Wenn Daten in einer organisierten Form erscheinen, wie z. B. in einer Datenbank, dann können diese als strukturiert angesehen werden (Mertens et al. 2023).

Vieweg et al. (2012) führen weiter aus, dass Daten verschiedene Datentypen haben können und somit z. B. als ganze Zahl (Integer), Datum (Date) oder als alphanummerische Zeichenkette (Char) auftreten. Diese Formatstrukturen werden als Datentypen bezeichnet (Mertens et

al. 2023). Die verschiedenen Datentypen werden in Variablen gespeichert, damit unterschiedliche Aktionen durchgeführt werden können (Abdel-Karim 2022). Laut Steyer (2024) ist der Vergleich von Datentypen für die weitere Operation essentiell, da verschiedene Datentypen nicht ohne Verständnis miteinander interagieren sollten. In der folgenden Tabelle werden die Standarddatentypen und die Basisdatentypen in Java nach Güting und Dieker (2018) dargestellt:

Tabelle 1: Standard Datentypen und Java Basisdatentypen (in Anlehnung an Güting und Dieker (2018), S. 42)

Standarddatentyp	**Java Basisdatentyp**
Integer	int (32 Bit)
	long (64 Bit)
	byte (8 Bit)
	short (16 Bit)
Real	float (32 Bit)
	double (64 Bit)
Boolean	boolean
Char	char

Wie in Tabelle 1 zu sehen ist, existieren verschiedene Java Basisdatentypen für einen Standarddatentyp. Laut Güting und Dieker (2018) unterscheiden sich je nach Programmierungssprache die Basisdatentypen. In diesem Fall werden die Basisdatentypen der Java-Programmiersprache thematisiert. Güting und Dieker (2018) führen weiter aus, dass sich die Datentypen Integer und Real bezüglich der Wertebereiche und bei der Genauigkeit der Darstellung unterscheiden. Eichstädt und Spieker (2024) erklären, dass die boolesche Variable im Gegensatz zu den anderen Basisdatentypen nur die Zustände wahr oder falsch annehmen kann. Die Gruppe der Standarddatentypen Real sind Fließkommazahlen und können nach der ganzen Zahl Nachkommastellen abbilden (Eichstädt und Spieker 2024). Die Gruppe der Standarddatentypen integer bilden hingegen nur Ganzzahlen ab. Zum Speichern von Buchstaben kann der Standarddatentyp char benutzt werden (Eichstädt und Spieker 2024). Laut Goll und Stamm (2024) sind diese Datentypen elementar und bilden die Grundlage für Zusammenstellungen. Weiterhin behaupten Cardelli und Wegner (1985), dass der Datentyp eine Spiegelung der Eigenschaften ist und unterschieden zwischen zwei Arten:

- Monomorphen Typsystem: Bei dem der Wert an einem Typ gebunden ist
- Polymorphen Typsystem: Mehrere Werte sind gebunden und gehören verschiedenen Typen an.

Aufgrund dieser Unterscheidung können Datentypen als komplex angesehen werden, die Listen, Mengen & Graphen speichern (Weiss 2013). Eine mögliche Speicherung von mehreren Variablen des gleichen Datentyps werden als Array bezeichnet (Güting und Dieker 2018). Aufgrund der verschiedenen Arten der Datentypen und den möglichen Operationen untereinander müssen die Datentypen anhand der Kriterien Ordnung und Rechnen kategorisiert werden (Cleve und Lämmel 2024). Cleve und Lämmel (2024) unterscheiden zwischen folgenden Kategorien:

- Nominal
- Ordinal
- Metrisch
    - Diskret
    - Kontinuierlich.

Unter nominalen Datentypen werden Daten verstanden, die keine Rangfolge besitzen und nur auf Gleichheit geprüft werden können (Cleve und Lämmel 2024). Ordinale Datentypen besitzen eine Ordnung und können anhand dieser verglichen werden (Cleve und Lämmel 2024). Cleve und Lämmel (2024) definieren metrische Datentypen als reelle Zahlen mit allen Ordnungsmerkmalen und erkennen das Rechnen untereinander an. Die Autoren unterteilen die metrischen Datentypen in diskret und kontinuierlich. Die diskreten Werte sind schrittweise größenveränderlich, wobei die kontinuierlichen Werte numerische Daten sind, die jede beliebige Zahl im Wertebereich annehmen können (Cleve und Lämmel 2024).

Ein weiterer Aspekt ist die Datenspeicherung, die bezüglich Daten und Datentypen beachtet werden muss. Die unverändert abrufbare Aufbewahrung über eine zeitliche Distanz, bedeutet Speichern und ist eine grundlegende Funktion in der Datenverarbeitung (Franz und Mattes 1991). Franz und Mattes (1991) führen weiter aus, dass zwischen einem internen und externen Speicher unterschieden wird. Schmidt (1993) behauptet, dass die externen Speicher hinzugezogen werden, wenn interne Speicherkapazitäten nicht mehr ausreichend sind. Mit der Datenspeicherung geht auch die Datenorganisation einher. Innerhalb der Datenorganisation wird der Zugriff auf die Daten und die Möglichkeit diese Daten schnell zu verändern thematisiert (Schmidt 1993). Weiterhin behauptet Schmidt (1993), dass die Daten sachlogisch strukturiert werden müssten, und führt die Hierarchie mit dem Datenfeld, dem Datensatz, der Datei und der Datenbank ein. Es folgen nach Mertens et al. (2023) die Definition der Begriffe:

- Datenfeld: Grundlegendes Element und besteht aus einem oder mehreren Zeichen
- Datensatz: Datenfelder, die logisch und inhaltlich zusammenhängen
- Datei: In einer Datei befinden sich alle zusammengehörende Datensätze, die das gleiche Format besitzen
- Datenbank: Sammlung von logisch zusammengehörenden Dateien.

Mertens et al. (2023) führen außerdem den Begriff des Datenbanksystems ein und behaupten, dass die aus einer Datenbank mit den zugehörigen Operationen, wie der Verwaltung bestehen müssten. Laut Vieweg et al. (2012) wird diese Verwaltung als Datenbankmanagementsystem definiert und bietet dem Administrator Werkzeuge wie Structured Query Language (SQL) an, um den schnellen Zugriff zu ermöglichen und dabei die Datensicherheit zu gewährleisten.

2.1.1 Datenebenen

Die im vorherigen Abschnitt eingeführte Hierarchie wird in diesem Unterabschnitt genauer erklärt und erweitert. Steiner (2014) unterteilt die Datenbanken in drei Hauptkategorien:

- Hierarchische Datenbanken
- Relationale und Objektrelationale Datenbanken
- Objektorientierte Datenbanken

In einer hierarchischen Datenbank besteht eine festgelegte Rangfolge für die Daten (Steiner 2014). Laut Steiner (2014) werden die Daten in der hierarchischen Datenbank in einer einzelnen sequentiellen Datei gespeichert und sind nicht flexibel bzgl. Veränderungen. Laut Vieweg et al. (2012) sind relational Datenbanken mehrere verschiedene Tabellen, die die Daten beinhalten und durch gleichartige Felder verbunden sind. Im Vergleich zu den hierarchischen Datenbanken ist die Reihenfolge nicht relevant, vielmehr die Eindeutigkeit der Daten (Vieweg et al. 2012). Um die Eindeutigkeit zu gewährleisten, werden Primärschlüssel benutzt (Mertens et al. 2023). Der Primärschlüssel definiert den eindeutigen Bereich des Datensatzes durch ein oder mehrere Felder (Vieweg et al. 2012). Im Gegensatz dazu existiert der Fremdschlüssel, der für die Relationen zwischen mehreren Tabellen verantwortlich ist (Vieweg et al. 2012). In den objektorientierten Datenbanken werden die Objekte, statt Tabellen fokussiert, die die Daten speichern (Steiner 2014). Die objektrelationalen Datenbanken verbinden Aspekte der relationalen und der objektorientierten Datenbanken (Steiner 2014). Laut Steiner (2014) können beliebige Datentypen verwendet werden und dabei kompatibel zu relational Datenbanken sein. Ein weiterer wichtiger Aspekt von Datenbanken ist die Speicherung der Daten. Es werden verschiedene Formate unterschieden, die den Gebrauch der Datenbanken beeinflussen (Vieweg et al. 2012). Vieweg et al. (2012) behaupten, dass das Format abhängig von der Menge der Daten sein kann und somit auch den lokalen bzw. den Onlinezugriff beeinflusst. Die Autoren unterscheiden zwischen den Text- und XML-, dateibasierten und serverbasierten Datenbankformat. Aufgrund der Thematik dieser Arbeit sind serverbasierte Datenbanken im Fokus, da diese mit riesigen Datenmengen in der Industrie genutzt werden (Vieweg et al. 2012). Der Zugriff auf diese Daten wird durch einen Server kontrolliert und durch Benutzernamen und Passwörter authentifiziert (Mielebacher 2024).

Die riesigen Datenmengen in der Digitalisierung haben zu einer Erweiterung der sachlogischen Struktur von Schmidt (1993) geführt. Laut Fasel und Meier (2016) ist der neue Trend Big Data geprägt durch das Internet-of-Things (IoT) sowie Web-Services. Die Datenmenge von Firmen der heutigen Zeit sind nicht zu vergleichen mit den Daten traditioneller Firmen (Fasel und Meier 2016). Big Data wird nach Badshah et al. (2024) durch die drei V´s definiert:

1. Volume: Bezieht sich auf die wachsende Datenmenge
2. Variety: Beschreibt die verschiedenen Datentypen
3. Velocity: Bezieht sich auf die Schnelligkeit des Zugriffs und der Entstehung der Daten.

Mertens et al. (2023) erweitern die Definition um zwei weitere Dimensionen:

4. Veracity: Betrifft die Genauigkeit bzw. Zuverlässigkeit der Daten, da diese aus unterschiedlichen Quellen stammen können

5. Value: Bezieht sich auf den Wert bzw. Nutzen der Daten für das Unternehmen.

Die Autoren erklären, dass Big Data für die Unternehmen neue Geschäftsmodelle eröffnet, da die Datenauswertung neue Möglichkeiten erhält. Laut García et al. (2015) geht mit der Datenauswertung in Big Data auch die Verwendung von neuen nützlichen Werkzeugen einher, die in Abbildung 1 dargestellt werden:

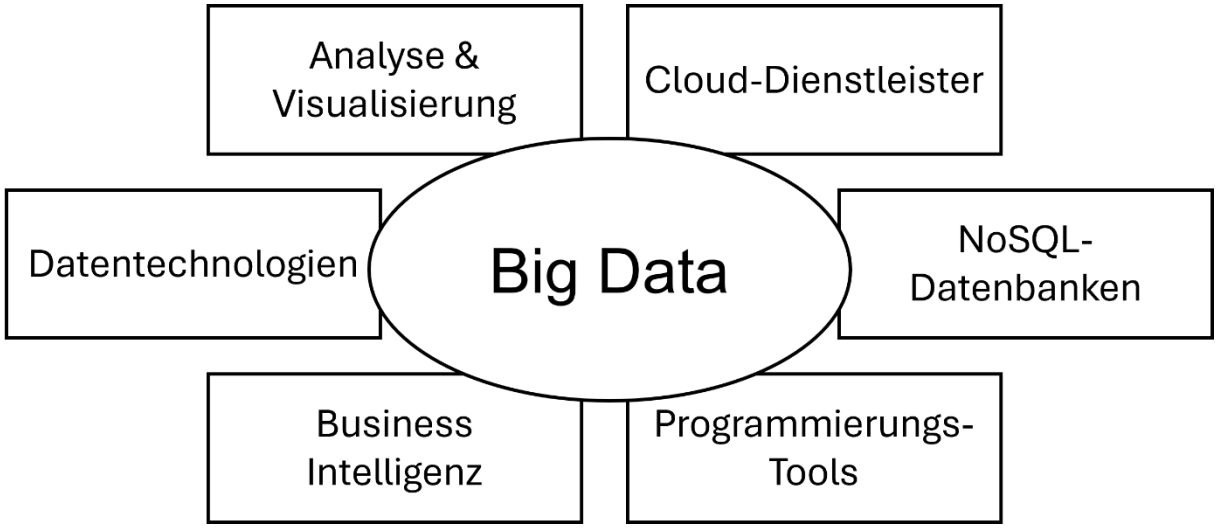


Abbildung 1: Big-Data-Werkzeuge (in Anlehnung an Badshah et al. (2024), S. 290)

Laut Badshah et al. (2024) existieren sechs zentrale Werkzeuge des Big-Data, um mit den enormen Datenmengen umzugehen, die in der Abbildung 1 zu sehen sind. Winkelhake (2024) führt weiter aus, dass je nach Fall die verschiedenen Werkzeuge unterschiedliche Methoden anwenden können. Winkelhake (2024) erweitert die Übersicht der Big-Data-Werkzeuge für das Verständnis und kategorisiert sie folgenderweise:

1. Datenintegration
2. Visualisierung
3. Analytische Verarbeitung
4. Daten-zugriff
5. Datengovernance und -sicherheit
6. Datenhaltung.

Weiterhin behauptet Winkelhake (2024), dass die modernen Big-Data-Werkzeuge sehr leistungsfähig bzgl. verschiedenen Datenformate sein und selbstständig Muster erkennen können. Es existieren verschiedene Anwendungsfälle für Big Data, wie z. B. in der Automobilindustrie (Gluchowski 2024). Gluchowski (2024) führt weiter aus, dass der Einsatz von Big Data die Produktionsplanung durch die Nutzung von Sensordaten verbessert. Laut Winkelhake (2024) existieren weitere Einsatzpotenziale, die in den verschiedensten Bereichen der Automobilindustrie umsetzbar sind. Dennoch läuft die Anwendung bzw. der Umstieg auf Big Data nicht reibungslos (Winkelhake 2024). Einer der Probleme ist die Skalierbarkeit, da veraltete Systeme nicht vollständig kompatibel mit Big Data sind (Badshah et al. 2024). Winkelhake (2024) führt weiter aus, dass nicht nur die Systeme, sondern auch die verschiedenen Organisationseinheiten im Unternehmen interoperabel funktionieren müssen. Ein weiterer wichtiger

Aspekt ist die Datenqualität, da bei Big Data hinterfragt werden muss, ob die enormen Datenmengen alle qualitativ auf dem gleichen Niveau sind (Badshah et al. 2024). Laut Demirbaga et al. (2024) werden dabei die Zuverlässigkeit und Genauigkeit der Daten thematisiert. Damit die Erkenntnisse aus Big Data nützlich sind, müssen die Daten verlässlich sein (Demirbaga et al. 2024).

2.1.2 Datenqualität

Zuletzt wurde die Bedeutung von Datenqualität in Big Data thematisiert. In diesem Unterabschnitt wird die generelle Datenqualität in den verschiedenen Datenebenen untersucht. Nach Hildebrand et al. (2021) wird die Datenqualität als Gesamtheit der Ausprägungen von Qualitätsmerkmalen eines Datenbestandes bzgl. dessen Eignung sowie festgelegte und vorausgesetzte Erfordernisse zu erfüllen definiert. Die Autoren führen weiter aus, dass die Qualität der Daten nur durch die Nützlichkeit beurteilt werden kann. Mertens et al. (2023) behaupten, dass Daten von schlechter Qualität seien, wenn sie Fehler, Duplikate, fehlende Werte oder falsche Formatierungen enthielten. Die Autoren führen weiter aus, dass aus diesen potenziellen Fehlerquellen wirtschaftliche Konsequenzen entstehen können. Hildebrand et al. (2021) führt die Datenqualitätspyramide für die Definition der Datenqualität ein und unterscheidet dabei zwischen den Qualitätsmerkmalen und der Qualitätsmetrik der Daten:

1. Datenqualität: Gesamtheit der Ausprägungen unterschiedlicher Qualitätsmerkmale
2. Datenqualitätsmerkmale: Einzelner Faktor zur Bewertung der Datenqualität
3. Datenqualitätsmetrik: Messbarer Faktor zur Bewertung der Datenqualitätsmerkmale.

Zur Strukturierung der Datenqualitätsmerkmalen und -metrik existieren verschiedene Ansätze (Hildebrand et al. 2021). Um ein besseres Verständnis bzgl. Datenqualität zu schaffen, werden die 15 Datenqualitätsdimensionen eingeführt (Strong et al. 1997). Es existieren noch weitere Definitionen von Datenqualitätsdimensionen, aber aufgrund von vielen Verweisen auf Strong et al. (1997), wurde diese Herangehensweise in Betracht gezogen. Nach Strong et al. (1997) werden diese 15 Dimensionen in vier Datenqualitätskategorien zugeordnet. Die folgende Tabelle veranschaulicht die Datenqualitätskategorien mit den verschiedenen Dimensionen:

Tabelle 2: Datenqualitätskategorien und -dimensionen (in Anlehnung an Strong et al. (1997), S. 104)

Intrinsisch	Zugänglichkeit	Kontextual	Repräsentativ
Genauigkeit	Verfügbarkeit	Relevanz	Interpretierbarkeit
Objektivität	Zugangssicherheit	Wert steigernd	Verständlichkeit
Glaubwürdigkeit		Aktualität	Prägnanz
Reputation		Vollständigkeit	Konsistenz
		Datenmenge	

Laut Hildebrand et al. (2021) wird unter der Kategorie Intrinsisch verstanden, dass die Informationen unabhängig von äußeren Faktoren so gegeben sind. Mertens et al. (2023) beschreiben die Datenqualitätsdimensionen aus der Kategorie Intrinsisch mit den Synonymen wie z. B. fehlerfrei und vertrauenswürdig. Die Kategorien Zugänglichkeit und Repräsentativ werden im Zusammenhang von IT-Systemen betrachtet (Hildebrand et al. 2021). Die Kategorie Zugänglichkeit thematisiert nicht nur die Datensicherheit in IT-Systemen, sondern auch die Schnelligkeit des Abrufens (Vogel-Heuser et al. 2024). Die Kategorie Repräsentativ befasst sich mit der Verständlichkeit und Einheitlichkeit der Daten (Mertens et al. 2023). Hildebrand et al. (2021) sieht die Dimensionen in der Kategorie Kontextual situationsbedingt an und findet diese für den Wettbewerb essenziell. Vogel-Heuser et al. (2024) betonen dabei, dass beachtet werden muss, welche der Dimensionen in mehreren Metriken gemessen werden können. Die Autoren führen weiter aus, dass nicht nur der Vergleich zwischen Daten und Datensätzen wichtig ist, sondern mit Werten gearbeitet werden soll. Heinrich et al. (2018) haben fünf Vorschriften zur Entwicklung von Metriken der Datenqualität eingeführt:

1. Minimal-/Maximalwert: die schlechteste und bestmögliche Datenqualität muss gekennzeichnet werden
2. Intervallskala: Differenzen zwischen Datensätze müssen innerhalb eines Intervalls eingeordnet werden können
3. Klarheit: Eindeutigkeit der Grenzen und Toleranzen
4. Aggregationsregeln: Auf einzelne Variablen, als auch auf ein ganzen Datensatz muss die Metrik anwendbar sein
5. Kosten: Dürfen bei der Bestimmung der Datenqualität nicht den Nutzen übersteigen (Vogel-Heuser et al. 2024; Heinrich et al. 2018).

Vogel-Heuser et al. (2024) führen weiter aus, dass für die Beurteilung der Datenqualität die oben genannten Vorschriften gelten sollten, um objektive, nützliche und wiederverwendbare Kenngrößen zu entwickeln. Weiterhin thematisieren die Autoren die Verbesserung der Datenqualität. Gegenmaßnahmen müssen getroffen werden, wenn Defizite in der Datenqualität ermittelt werden (Vogel-Heuser et al. 2024). Es existieren verschiedene Werkzeuge, die die Datenqualität überprüfen und die Defizite ausgleichen (Baumöl et al. 2000). Lösungsansätze für Defizite von bestehenden Daten können technisch, designbezogen, organisatorisch und architektonisch sein (Baumöl et al. 2000). Organisatorische und architektonische Lösungsansätze befassen sich z. B. mit Zugänglichkeitsdefiziten und nicht interpretierbaren Schlüsselsystemen (Baumöl et al. 2000). Laut Vogel-Heuser et al. (2024) existieren auch einzelne inkonsistente Fehler, die mühelos entfernt werden können. Die Autoren erklären, dass die Maßnahmen für gute Datenqualität nicht nur auf bestehende Daten angewendet werden können, sondern auch auf neue Daten. Dabei ist es von Bedeutung, die Prozesse und Analyse anzupassen und somit die Entstehung der Fehler bzw. schlechter Datenqualität zu verhindern (Vogel-Heuser et al. 2024). Die Vorbereitung der Suche von neuen Daten muss aufgrund der Datenqualität also angepasst werden. Mit dem Verständnis von qualitativ hochwertigen Daten werden die neuen Daten bzw. neue Datenzusammenhänge besser bewertet. Die Suche bzw. das Erkennen von nützlichen Daten werden im Data Mining dargelegt.

2.2 Data Mining

Das Data Mining ist aufgrund der Digitalisierung und der vorhandenen Datenmenge ein relevantes Thema (García et al. 2015). Die wachsende Menge an Daten bedeutet nicht, dass diese Daten sinnvoll genutzt werden bzw. einen Mehrwert erbringen (García et al. 2015). Im Data Mining wird dieser Aspekt behandelt, da es sich mit der Extraktion implizit vorhandenem, einfachem und nützlichem Wissen aus verschiedenen aufgebauten Datenbanken befasst (Bissantz und Hagedorn 2009). Das Data Mining steht für die Mustererkennung und erkennt im Prozess Beziehungen zwischen den Daten (Bissantz und Hagedorn 2009). Außerdem wird das Data Mining mithilfe von statistischen Methoden und maschinellem Lernen auch als Prozess des automatischen Gewinnens von potenziell nützlichem Wissen verstanden (Vogel-Heuser et al. 2024). Interessante Muster in den Daten werden in der Business Analyse zu Prognosen über zukünftige Geschehnisse benutzt (Gluchowski et al. 2021; D'Onofrio und Meier 2021). Cleve und Lämmel (2024) betrachten nicht nur die Anwendungen der verschiedenen Methoden als Data Mining, sondern auch die verschiedenen Phasen davor und danach. Das bedeutet, dass das Selektieren, die Datenvorbereitung und die Evaluation zum Data-Mining-Prozess zählen (Cleve und Lämmel 2024). Nach Cleve und Lämmel (2024) werden diese drei bekannten Modelle für Data-Mining-Prozesse unterschieden:

1. Knowledge Discovery in Databases (KDD) nach Fayyad et al. (1996)
2. Cross-Industry-Standard-Process-for-Data-Mining (CRISP-DM)
3. Sample, Explore, Modify, Model and Assess (SEMMA).

Laut Runkler (2015) umfasst das allgemeine KDD mit Selektieren, Vorverarbeitung, Transformation, Data-Mining und die Evaluation sechs Phasen. Der Autor führt weiter aus, dass das CRISP-DM mit Domänenverständnis, Datenverständnis, Datenaufbereitung, Modellierung, Evaluierung und Einsatz ebenfalls sechs Phasen umfasst. Nach Cleve und Lämmel (2024) besteht SEMMA aus fünf Phasen und beinhaltet Sammeln von Daten, Datenuntersuchung, Modifizieren, Modellieren und Evaluieren. Die Autoren behaupten, dass das wahre Data Mining als Modellieren ein Teilschritt innerhalb der verschiedenen Prozesse sei. Han et al. (2011) verdeutlichen die Iteration innerhalb der Modelle wie z. B. im KDD. Um die Phasen genauer zu thematisieren, wird das vereinfachte Modell von Runkler (2015) benutzt, bei dem die Gemeinsamkeiten von den verschiedenen Data-Mining-Prozessen in der folgenden Tabelle dargestellt werden:

Tabelle 3: Data-Mining-Prozesse (in Anlehnung an Runkler (2015), S. 3)

Vorbereitung	Vorverarbeitung	Analyse	Nachbereitung
Planung	Säuberung	Visualisierung	Interpretation
Datensammlung	Filterung	Korrelation	Dokumentation
Merkmalsgenerierung	Ergänzung	Regression	Auswertung
Datenauswahl	Korrektur	Prognose	

Vorbereitung	Vorverarbeitung	Analyse	Nachbereitung
	Standardisierung	Klassifikation	
	Transformation	Clustering	

Wie zu erkennen ist, werden die Phasen Vorbereitung und Vorverarbeitung vor dem tatsächlichen Data Mining ausgeführt und bilden die Grundlage für die Analyse (Runkler 2015). In den ersten Phasen des Data-Mining-Prozesses wird das Verständnis der zu sammelnden Daten angeschnitten (Olson et al. 2024). Dabei ist es von Bedeutung, dass die Ziele festgelegt und die Daten so ausgewählt werden, damit die einen Mehrwert generieren (Olson et al. 2024). Die ausgewählten Daten werden auf ihre Qualität und auf die Beziehungen untereinander untersucht (Cleve und Lämmel 2024). Die Vorbereitung wird mit einem Projektplan abgeschlossen (Cleve und Lämmel 2024). Die zweite Phase nach dem vereinfachten Modell von Runkler (2015) geht auf die Datenvorverarbeitung bzw. Vorbereitung ein. In diesem Kontext werden die gesammelten Daten gesäubert und fehlerhafte oder inkonsistente Daten werden korrigiert (Cleve und Lämmel 2024). Die Datenvorverarbeitung wird im Abschnitt 2.4 genauer zur Sprache gebracht. Die dritte Phase befasst sich mit der Analyse und dem wahren Data-Mining-Prozess. Laut Olson et al. (2024) wird dabei eine Data-Mining-Software angewendet, um Ergebnisse für die Gegebenheiten zu generieren. Die Autoren behaupten, dass verschiedene Methoden eingesetzt werden können. Nach Olson et al. (2024) wird der vorhandene Datensatz in einen Trainings- und einen Testsatz aufgeteilt, wobei ersteres das Modell entwickelt und letzteres für die Überprüfung zuständig ist. Die Phase der Analyse und Anwendung des Data-Mining wird im nächsten Abschnitt genauer untersucht. Schließlich folgt nach der Analyse die letzte Phase des vereinfachten Modells nach Runkler (2015), und zwar die Nachbereitung. In dieser Phase werden die Resultate geprüft und bewertet (Cleve und Lämmel 2024). Die vordefinierten Ziele aus der ersten Phase werden geprüft und Ergebnisse interpretiert (Cleve und Lämmel 2024).

Aus den Phasen des Data-Mining-Prozesses resultiert, dass auch Herausforderungen betrachtet werden müssen (Chen et al. 1996). Laut Chen et al. (1996) existieren folgende Herausforderungen im Data Mining:

- Umgang mit verschiedenen Datentypen
- Effizienz und Skalierbarkeit der Data-Mining-Methoden
- Nutzbarkeit und Aussagekraft von den Ergebnissen aus dem Data Mining
- Verschiedene Anfragen und Ergebnisse über mehreren Abstraktionsebenen des Data Mining ausdrücken
- Datensicherheit
- Informationen aus verschiedenen Quellen verwalten.

Um die Herausforderungen besser zu verstehen, werden im nächsten Abschnitt die Methoden des Data Mining vorgestellt und in den Kontext des gesamten Prozesses eingeordnet.

2.3 Methoden des Data Mining

In diesem Abschnitt werden die verschiedenen Methoden des Data Mining behandelt, die in der Phase der Analyse benutzt werden können. Dafür wird für jede Kategorie der Methode ein Verfahren genauer thematisiert. In Abschnitt 2.2 wurde die Analyse als Phase des Data-Mining-Prozesses vorgestellt. Für die Analyse der Daten werden wie vorher genannt, verschiedene Methoden angewendet (Olson et al. 2024). Laut Gluchowski et al. (2021) werden die Methoden anhand ihres Einsatzzweckes unterschieden. Die Autoren führen weiter aus, dass für Vorhersagen Klassifikations- oder Regressionsmethoden und für Beschreibungen Clustering oder Assoziationsanalysen benutzt werden. Weiterhin behaupten die Autoren, dass die Vorhersageverfahren überwacht und die Beschreibungsverfahren unüberwacht seien.

Die Klassifikation gehört zu Datenanalyse, in der Modelle erstellt und Daten zugeordnet werden können (Han et al. 2011). Wie im allgemeinen Vorgehen erläutert, besteht das Modell der Klassifikation aus einem Lernschritt und Klassifizierungsschritt (Han et al. 2011). Zuerst wird das Modell mithilfe von vorbestimmten Daten trainiert (Han et al. 2011). Das Training hat den Sinn, dass das Modell mithilfe einer Gegebenheit oder Funktion die Daten separiert (Han et al. 2011). Han et al. (2011) behaupten, dass die Klassifikation dann im zweiten Schritt durchgeführt werden kann, wenn die Klassifikationsgenauigkeit mit dem Trainingsdatensatz passt. Laut Gluchowski et al. (2021) eignen sich Entscheidungsbäume als Klassifikationsmethode. Das Modell benötigt in der Entwicklung einen historischen Datenbestand der anhand eines Merkmals klar zugeordnete Daten beinhaltet (Gluchowski et al. 2021). Der Entscheidungsbaum basiert auf Regeln, die durch bekannte Ergebnisse des Datenbestands erzeugt werden können (Gluchowski et al. 2021). Aggarwal (2017) sieht den Entscheidungsbaum als eine hierarchische Klassifikation und bezeichnet die Knoten im Verfahren als Teilungskriterium. Das bedeutet in jeder Ebene wird anhand der Merkmalsvariablen eine vorher definierte Entscheidung getroffen, aus der dann eine Teilung der Daten resultiert (Aggarwal 2017). Jeder Knoten bzw. Entscheidung ist nicht unbedingt nur an eine Variable gebunden (Aggarwal 2017). Dies bedeutet, dass auch mehrere Merkmalsvariablen in einer Ebene bzw. an einem Knoten eine Entscheidung beeinflussen (Aggarwal 2017). Laut Gluchowski et al. (2021) wird an jedem Knoten ein Homogenitätsmaß gemessen. Dieses Maß gibt den Grad der Aufteilung an (Gluchowski et al. 2021). Die Homogenität wird als rein bezeichnet, wenn die Bedingung maximal erfüllt wird (Gluchowski et al. 2021). Nach Aggarwal (2017) existieren verschiedene Methoden, um den Teilungsgrad von Merkmalvariablen zu messen. Mithilfe des Gini-Index wird die Reinheit an einem Knoten gemessen, wodurch die Auswahl der Merkmalsvariable bewertet werden kann (Aggarwal 2017). Der Gini-Index nach Aggarwal (2017) sieht folgenderweise aus:

$$G(vi) = 1 - \sum_{j=1}^{k} p_j^2$$

Formel 1: Gini-Index

Der Wert von G ist zwischen 0 (maximale Reinheit) und 1 (Aggarwal 2017). Das p definiert die Anteile der Merkmalsvariablen an einem Knoten und k steht für die Anzahl der Klassen (Aggarwal 2017). Eine weitere Methode zur Beurteilung der Reinheit an einem Knoten ist die Entropie E (Gluchowski et al. 2021). Die Formel der Entropie nach Gluchowski et al. (2021) lautet:

$$E(T) = -\sum_{i=1}^{k} p_i * log_2 p_i$$

Formel 2: Entropie

Das p_i steht für die relativen Häufigkeiten und das k wieder für Anzahl der Klassen (Gluchowski et al. 2021). Homogenität wird gewährleistet, wenn E den Wert 0 annimmt (Gluchowski et al. 2021). Das bedeutet, dass bei einem hohen Wert der Entropie eine Unordnung der Klassen vorhanden ist (Aggarwal 2017). Gluchowski et al. (2021) behauptet, dass aus der Ausgangslage der Entropie bzw. der Berechnung der Homogenität, eine weitere Unterteilung des Knoten in Teilknoten erfolgen könnte. Mit der folgenden Formel nach (Gluchowski et al. 2021) werden mögliche Teilknoten berechnet:

$$IG(T, A) = E(T) - \sum_{a \in A} \frac{|T_a|}{|T|} * E(T_a)$$

Formel 3: Informationsgewinn an den Teilknoten

Aus dieser Formel resultiert der Informationsgewinn IG (Gluchowski et al. 2021). Ergänzend zur Berechnung der normalen Entropie wird hierbei das Verhältnis der relativen Größe der Knoten T_a zum Ausgangsknoten T gemessen (Gluchowski et al. 2021). Mit den Berechnungen der Entropie und der Informationsgewinne können Entscheidungsbäume entstehen, die mehrere homogene Knoten besitzen (Gluchowski et al. 2021). Durch Prepruning kann vor der Erzeugung des Entscheidungsbaums die Anzahl der Knotenebenen vorgegeben werden (Gluchowski et al. 2021). Dies ist notwendig, um Overfitting zu vermeiden (Han et al. 2011). Laut (Han et al. 2011) wird bei dem Prepruning Abschnitte(Äste) des Entscheidungsbaum ausgeschnitten, d. h. am Knoten wird nicht weiter geteilt, da die folgenden Entscheidungen nicht mehr relevant sind. Außerdem ist ein Postpruning möglich, um im Nachhinein den Entscheidungsbaum zu begrenzen (Gluchowski et al. 2021). Laut (Olson et al. 2024) existiert die Möglichkeit mehrere ungeschnittene Entscheidungsbäume in einem zu betrachten. Diese werden als Random Forest bezeichnet und finden für große Datensätze Anwendung (Olson et al. 2024). Laut (Aggarwal 2017) sind Entscheidungsbäume einfach zu implementieren und erzeugen hohe Interpretationsmöglichkeiten. Weitere bekannte Verfahren zu den Klassifikationen werden von Olson et al. (2024) genannt:

- Support-Vector-Machines
- Neural Networks.

Nachdem ein Verfahren der Klassifikation vorgestellt wurde, wird jetzt die Regression thematisiert. Durch eine oder mehrere unabhängigen Variablen wird die beobachtete und abhängige

Variable in der Regression erklärt (Gluchowski et al. 2021). Also werden Beziehungen zwischen Variablen analysiert (Backhaus et al. 2023). Die Modellierung der Regression ist ein Vorhersageprozess (Aggarwal 2017). Die lineare Regression ist die einfachste Version für lineare Zusammenhänge zwischen Merkmalen (Runkler 2015). Die Abhängigkeit jeder Zielvariablen y_i von den zugehörigen unabhängigen Variablen X_i wird in folgenderweise nach Aggarwal (2017) modelliert:

$$y_i \approx W * X_i \quad \forall i \in \{1, \dots, n\}$$

Formel 4: Abhängigkeit der Variablen in der Regression

Das W steht in diesem Kontext für den Koeffizientenvektor, der mithilfe des Trainingssatzes verbessert werden muss, damit die Fehlerrate des Models sinkt (Aggarwal 2017). Laut Runkler (2015) und Aggarwal (2017) wird für die lineare Regression der quadratische Regressionsfehler minimiert:

$$O = \sum_{k=1}^{n} (W * X_i - y_i)^2 = ||DW^T - y||^2$$

Formel 5: Quadratische Regressionsfehler Minimierung

Das D ist eine n x d Matrix, in der der i-te Datenpunkt den Vektor für X_i bildet (Aggarwal 2017). Runkler (2015) führt weiter aus, dass aus den Mittelwerten und der Kovarianzmatrix von X die Parameter aller paarweisen linearen Regressionsmodelle für einen Datensatz bestimmt werden können. Durch die Berechnungen der notwendigen Bedingungen bzw. der Extrema resultieren die Werte der Vorfaktoren zum Aufstellen der Regression (Runkler 2015). Die Regressionsfunktion muss mithilfe des gegebenen Datensatzes abgeschätzt werden (Backhaus et al. 2023). Der Schwerpunkt der Daten bestimmt den Verlauf der Regressionsgerade (Backhaus et al. 2023). Für das Schätzen vom Modell werden die Parameter mithilfe der Kovarianz und der Varianz berechnet (Backhaus et al. 2023). Danach erfolgt die Analyse der Residuen, bei der sie zuerst berechnet und danach auf ihre Streuung überprüft werden (Backhaus et al. 2023). Die Güte des Modell wird mithilfe des Bestimmtheitsmaß validiert, sodass danach eine Vorhersage von neuen Werten getätigt werden kann (Backhaus et al. 2023). Falls mehrere unabhängige Variablen existieren, kann die lineare Regression in eine multiple Regression erweitert werden, wodurch mehrere Regressionsgeraden nicht notwendig sind. (Backhaus et al. 2023).

Nachdem die Regressionsanalyse behandelt wurde, werden jetzt die deskriptiven Methoden des Data Mining erklärt. Eine der Methoden ist das Clustering, bei der Datenobjekte gruppiert werden (Gluchowski et al. 2021). Unmarkierte Daten werden Clustern zugeordnet (Runkler 2015). Laut Aggarwal (2017) existieren verschiedene Szenarien des Clustering, aufgrund unterschiedlicher Datentypen. Der Autor führt weiter aus, dass die Dimension der Daten ebenfalls von Bedeutung ist, da eine hohe Dimension den Aufwand erhöht und die Darstellung erschwert. Im Prozess des Clustering werden Klassenzugehörigkeiten der Daten beachtet, sodass die Cluster möglicherweise den Klassenzugehörigkeiten entsprechen (Runkler 2015). Cleve und Lämmel (2024) ergänzen die Definition und erklären, dass eine gegebene Menge

von Daten in verschiedene Teilmengen zerlegt werden. Die Datenobjekte von verschiedenen Clustern müssen klar voneinander unterscheidbar sein, wohingegen Objekte innerhalb desselben Clusters sich ähneln dürfen (Cleve und Lämmel 2024). Das Clustering wird nach Aggarwal (2017) auch als eine Zusammenfassung von mehreren Datenobjekten betrachtet. Für die Erzeugung von Clustern können mehrere mathematische Methoden benutzt werden (Runkler 2015). Aggarwal (2017) ergänzt, dass verschiedene Modelle für die Cluster-Analyse existieren, und dass diese abhängig von unterschiedlichen Faktoren wie z. B. Datentypen sind. Die generelle Funktionsweise des Clustering basiert auf Distanz- oder Abstandsfunktionen (Cleve und Lämmel 2024). Vor dem Clustering muss eine Merkmalsauswahl durchgeführt werden, damit Ausreißer entfernt werden, die sich nicht gut Clustern lassen (Aggarwal 2017). Für die Merkmalsauswahl können zwei Klassen von Modellen verwendet werden (Aggarwal 2017). In den Filter-Modellen werden jedem Merkmal eine Punktzahl anhand eines Ähnlichkeitskriterium vergeben (Aggarwal 2017). Die Wrapper-Modelle sind die zweite Klasse und dabei wird ein iterativer Clustering-Algorithmus verwendet, der die Qualität der Datenobjekte mithilfe von Teilmengen bewertet (Aggarwal 2017). Aggarwal (2017) behauptet, dass zur Merkmalsauswahl eine kombinierte Version von beiden Modellklassen eingesetzt werden sollte. Für den Clustering-Algorithmus stehen verschiedene Verfahren zur Auswahl. Der k-Means-Algorithmus wird in diesem Fall erklärt und ist in der Abbildung 2 zu erkennen. Laut Cleve und Lämmel (2024) ist dieser Algorithmus ein partitionierendes Verfahren der Clusteranalyse.

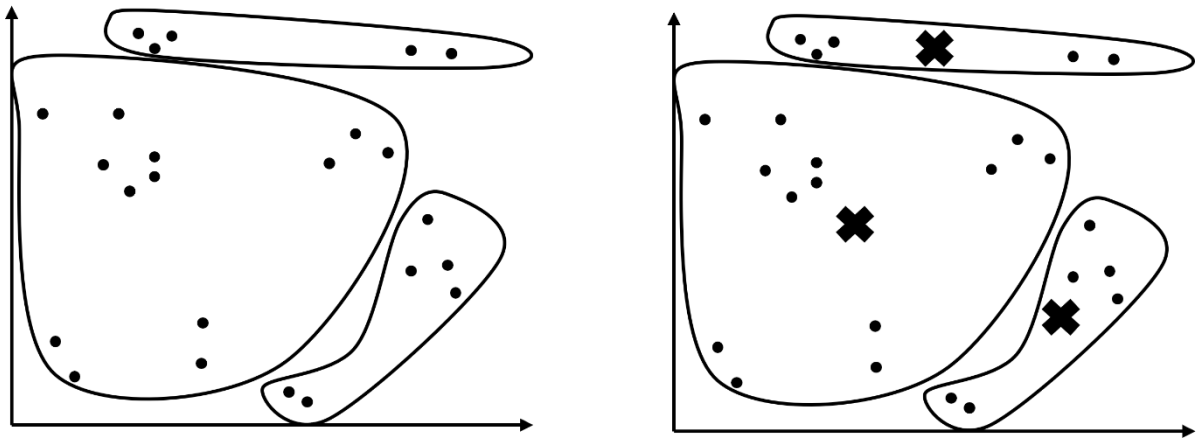


Abbildung 2: Initiale Cluster und Bilden der Centroiden (in Anlehnung an Cleve und Lämmel (2024), S. 158)

Partitionierende Verfahren kennzeichnet aus, dass die Cluster sich nicht überlappen (Dehmer 2005). Im k-Means-Algorithmus wird dazu die Anzahl der gesuchten Cluster vorgegeben (Cleve und Lämmel 2024). Wie auf der linken Seite in Abbildung 2 zu sehen ist, werden zuerst initiale Cluster gebildet, die durch Iterationen immer wieder neu definiert werden. Auf der rechten Seite der Abbildung werden die Centroide (Schwerpunkte) dargestellt, die den Cluster repräsentieren (Cleve und Lämmel 2024). Nach den ersten Centroiden werden die Abstände von den Punkten zu den Centroiden berechnet und neue Cluster erstellt (Cleve und Lämmel 2024). Die neuen Cluster sind in Abbildung 3 zu sehen. Diese basieren auf Berechnungen und kriegen durch die Distanzkalkulationen neue Ordnungen (Cleve und Lämmel 2024).

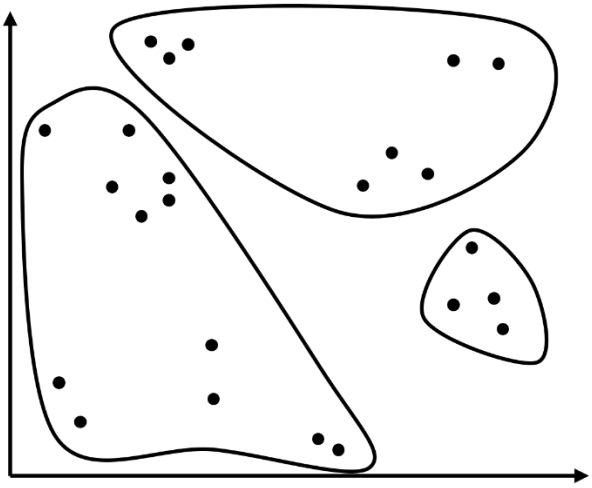


Abbildung 3: Neu berechnetete Centroiden (in Anlehnung an Cleve und Lämmel (2024), S. 158)

Laut Gluchowski et al. (2021) wird der iterative Prozess durchgeführt, bis keine Verbesserung zu erkennen ist. Die Autoren führen weiter aus, dass durch eine Summierung der quadrierten euklidischen Distanzen der Objekte zum Centroid berechnet werden. Die Euklidische Distanz für zwei Datensätze x und y mit den Attributausprägungen x = (x_1, x_2, …, x_n) wird folgenderweise nach Gluchowski et al. (2021) berechnet:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

Formel 6: Euklidische Distanzberechnung

Nach den Berechnungen fehlt der Algorithmus, der die Datenobjekte in Cluster überführt (Gluchowski et al. 2021). Laut Cleve und Lämmel (2024)wird im Algorithmus die Distanz zu den Centroiden berechnet und basierend auf der geringsten Distanz zugeordnet. Dieser Prozess wird mit einer If-Bedingung geprüft und durch eine Boolesche Ausgabe bestätigt (Cleve und Lämmel 2024). Die Vorteile des Verfahrens sind die einfache Implementierung, und dass nicht so viele Iterationen notwendig sind wie erwartet (Cleve und Lämmel 2024). Zu den Nachteilen gehört der erste Schritt, indem man initiale Cluster bilden muss und dabei durch Ausreißer getäuscht werden kann (Cleve und Lämmel 2024).

Die letzte Methode des deskriptiven Data Mining ist die Assoziationsanalyse. In dieser werden Korrelationen zwischen gemeinsam auftretenden Objekten gesucht (Cleve und Lämmel 2024). Ziel der Assoziationsanalyse ist es Regeln aufzustellen, die Zusammenhänge zwischen den Datenobjekten aufweisen (Cleve und Lämmel 2024). Laut Aggarwal (2017) ist die Assoziationsanalyse mit viel Aufwand verbunden, da uninteressante oder überflüssige Daten ebenfalls analysiert werden. Nachfolgend wird der Apriori Algorithmus vorgestellt, der ein Standardverfahren für die Assoziationsanalyse ist (Cleve und Lämmel 2024). Zuerst wird der Frequent Itemset definiert, der eine Gruppe von Objekten ist, die mit einer bestimmten Wahrscheinlichkeit zusammen auftreten (Aggarwal 2017). Der Frequent Itemset muss den Schwellenwert erreichen, sodass dieser benutzt werden kann (Cleve und Lämmel 2024). Laut Cleve und Lämmel (2024) wird der Algorithmus mit diesen zwei Schritten durchgeführt:

1. Frequent Itemsets ermitteln

2. Assoziationsregeln bestimmen.

Durch die Ermittlung der Frequent Itemsets wird der Suchbereich innerhalb der Daten reduziert und somit auch der Aufwand (Cleve und Lämmel 2024). Mithilfe einer Iteration werden potenzielle Objekte geprüft, die zu den Frequent Itemsets gehören könnten und bei nicht Übereinstimmung entfernt, da deren Teilmengen Frequent sein müssen (Cleve und Lämmel 2024). Nachdem die Frequent Itemsets bestimmt wurden, werden diese in zwei nicht leere Mengen aufgeteilt und die Konfidenz berechnet (Cleve und Lämmel 2024). Die Konfidenz gibt die Wahrscheinlichkeit des Zusammenhangs zwischen den Objekten an (Cleve und Lämmel 2024). Die Assoziationsregeln wurden erfolgreich aufgestellt, wenn der Schwellenwert der Frequent Itemsets und der minimale Konfidenzwert erreicht wurde, wobei der Konfidenzwert über den Schwellenwert der Frequent Itemsets definiert ist (Cleve und Lämmel 2024). Dies waren die Analysemethoden des Data-Mining, die je nach Anwendungsfall und Ziel innerhalb des Data-Minig-Prozesses benutzt werden können.

2.4 Datentransformation in der Datenvorverarbeitung

Wie im vorherigen Prozess des Data Mining erklärt wurde, ist eine Datenvorverarbeitung notwendig. In diesem Abschnitt wird der Fokus auf die Datentransformation gelegt, die als Datenvorverarbeitungsschritt angewendet wird. Vor der Thematisierung der Datentransformation im Detail, wird die Relevanz der Datenvorverarbeitung verdeutlicht. Laut Runkler (2015) müssen Daten skaliert und transformiert werden, da sie im Realen meistens Fehler enthalten. Cleve und Lämmel (2024) ergänzen zur Definition, dass nicht nur Fehler thematisiert werden, sondern generell Daten, die als ungeeignet klassifiziert werden. Die Autoren sprechen in diesem Sinne von Qualität des Zieldatenbestands und behaupten, dass die Daten durch geeignete Verfahren angepasst werden müssen. Die Datenvorverarbeitung soll invariante Datendarstellungsformen gewährleisten und Wertebereiche eingrenzen, um die Bearbeitung von Daten zu vereinfachen Cleve und Lämmel (2024). Die zuvor als ungeeignet definierten Daten können nicht alle gleich betrachtet werden (Cleve und Lämmel 2024). Im weiteren Verlauf des Data Mining können die als ungeeigneten Daten auf Mängel bzw. Probleme hinweisen, die in der ersten Betrachtung nicht direkt aufgefallen sind (Cleve und Lämmel 2024). Ein Beispiel dafür sind Ausreißer, die aufgrund einer erkennbaren Abweichung zum restlichen Datenbestand nicht ignoriert und direkt gelöscht werden können (Cleve und Lämmel 2024). Die Datenvorverarbeitung besteht aus mehreren Schritten und unterscheiden sich je nach Autor. Laut Jassim und Abdulwahid (2021) sind die wichtigsten Schritte der Datenvorverarbeitung:

- Datenreinigung
- Datenintegration und -sammlung
- Datentransformation.

In der Datenreinigung werden fehlerhafte, falsch formatierte und Duplikate korrigiert oder aus dem Datensatz entfernt (Jassim und Abdulwahid 2021). Die Datenintegration thematisiert das Verknüpfen und Sammeln von Daten aus mehreren Quellen und beachtet dabei die sinnvolle Nutzung bzw. den Vergleich dieser Daten (Jassim und Abdulwahid 2021). In der Datentransformation werden die Wertebereiche von Merkmalen angepasst (Runkler 2015). Cleve und

Lämmel (2024) führen aus, dass die Datentransformation kurz vor dem eigentlichen Data-Mining-Schritt, der Datenanalyse benutzt wird. Die Autoren setzen die Definition von Runkler (2015) fort und behaupten, dass die Datentransformation Daten in eine Form verwandeln, mit der im Data Mining gearbeitet werden kann. Jassim und Abdulwahid (2021) behaupten, dass die Datentransformation für das generelle Datenmanagement und für die Datenintegration von Bedeutung sei. Cleve und Lämmel (2024) beziehen sich bei der Datenvorverarbeitung auf die Qualitätsmerkmale und betrachten diese als Kriterien zur Einschätzung. Die Autoren führen weiter aus, dass Transformationen existieren, die verfahrensabhängig und -unabhängig genutzt werden können. In diesem Kontext wird die Weiterverarbeitung bzw. die Analyse des Data Mining beachtet (Cleve und Lämmel 2024). Weiterhin geben Cleve und Lämmel (2024) Beispiele an für die Anpassungen erforderlich sein können:

- Datentypen
- Konvertierungen oder Codierungen
- Zeichenketten
- Datumsangaben
- Maßeinheiten und Skalierungen.

Laut Aggarwal (2017) ist das Transformieren von Datentypen das relevanteste Element im Data-Mining-Prozess. Der Autor führt weiter aus, dass je nach Fall Datensätze verschiedene Typen beinhalten können. Die Vermischung der Datentypen erschwert die Bearbeitung von Analysten und kann bei der Umwandlung zu Verlusten von der Datengenauigkeit führen (Aggarwal 2017). Aufgrund des Data Mining ist zu erkennen, dass numerische Datentypen am häufigsten auftreten und diese fokussiert werden (Aggarwal 2017). Außerdem existieren Szenarien, in denen einige Datentypen in Graphen transformiert werden (Aggarwal 2017). Um diese Datentypumwandlungen genauer zu betrachten, werden im kommenden Unterabschnitt die verschiedenen Datentransformationsverfahren vorgestellt.

2.4.1 Datentransformationsverfahren

Es existieren verschiedene Verfahren, um Daten zu transformieren. Es folgen die verfahrensunabhängigen Transformationen nach Cleve und Lämmel (2024):

- Kombination oder Separierung von Attributen
- Berechnung abgeleiteter Werte
- Aggregation
- Glättung.

Diese Verfahren sind nur für Datenbestände relevant, die in sich konsistent sein sollen (Cleve und Lämmel 2024).

Cleve und Lämmel (2024) thematisieren Anpassungen, wie z. B. Codierungen und stellen das binäre Codieren als eine mögliche Methode vor. Eine Form des Codieren ist das One-Hot-Encoding (OHE) (Cleve und Lämmel 2024). Aggarwal (2017) beschreibt dieses Verfahren als Binarization. Laut Kim (2024) zählt das OHE zum Feature Engineering und codiert kategoriale Datentypen in ein binäres Format. Der Autor erklärt, dass jede Kategorie durch eine binäre

Variable dargestellt werden kann. Für diese Methode wird eine Matrix erstellt, wo die Spalten nach der jeweiligen Kategorie benannt werden (Fan et al. 2021). Bei einer Übereinstimmung werden die Kategorien mit einer 1 gekennzeichnet (Kim 2024). Wenn die Kategorie nicht passt, wird der Variable eine 0 zugeordnet (Kim 2024). Durch eine simple Nummerierung von 1 startend, könnte es wie eine ordinale Fortführung wirken (Cleve und Lämmel 2024). Durch die binäre Methodik im OHE können also Daten in eine numerische Form transformiert werden, ohne eine ordinale Struktur zu erzeugen (Cleve und Lämmel 2024).

Laut Cleve und Lämmel (2024) ist eine weitere Art der Datentransformation die Kombination von Attributen. Han et al. (2011) beschreiben als mögliches Verfahren die Generalisierung. In der Generalisierung werden nominale Attribute betrachtet, die endliche Werte ohne eine Rangfolge besitzen (Han et al. 2011). In diesem Verfahren werden detaillierte Attribute durch höhere Ebenen repräsentiert (Alasadi und Bhaya 2017). Als Beispiel könnte man die Attribute Straßen durch die höhere Ebene Stadt generalisieren (Alasadi und Bhaya 2017). Laut Jiang et al. (1999) ist bei der Generalisierung Domänenwissen von Bedeutung, da die Repräsentation nicht immer automatisch erfolgt. Weiterhin muss bei der Generalisierung beachtet werden, dass diese zur Informationsverlusten führen kann, da die detaillierten Attribute nur durch höhere Ebenen dargestellt werden (Jiang et al. 1999). Han et al. (2011) unterteilen die Generalisierung in vier Gruppen:

1. Spezifizierung einer partiellen Ordnung der Attribute anhand eines Schemas durch einen Experten: Mithilfe eines vorgegebenen Schemas, wie z. B. durch eine relationale Datenbank, kann der Experte die Generalisierung durchführen.
2. Spezifizierung eines Teils einer Hierarchie durch explizite Datengruppierung: Statt eine gesamte Hierarchie zu schaffen, werden in dieser Methode mehrere Teilhierarchien gebildet.
3. Spezifizierung von Attributen, aber nicht durch die partielle Ordnung: In diesem Fall wird keine partielle Ordnung vom Anwender benutzt, sondern die Anzahl der Attribute. Die Autoren behaupten, dass die Anzahl der Attribute in höheren Hierarchien geringer seien und deshalb automatisch zugeordnet werden können.
4. Spezifizierung einer Teilmenge von Attributen: Mithilfe der Datensemantik innerhalb einer Datenbank soll eine kleine Teilmenge der Attribute zu einer automatischen Abrufung von weiteren relevanten Attributen führen, die mit der Teilmenge verknüpft sind.

Nachdem die Generalisierung und die verschiedenen Gruppen vorgestellt wurden, werden jetzt die Berechnung von abgeleiteten Werten nach Cleve und Lämmel (2024) aufgegriffen. Cleve und Lämmel (2024) betrachten simple mathematische Operationen, bei denen neue Attribute entstehen, als Methode der Datentransformation. Als Methode für die Berechnung von solchen abgeleiteten Werten stellen die Autoren Differenzberechnungen vor (Cleve und Lämmel 2024). Ein Beispiel hierfür ist die simple Berechnung vom Gewinn, durch die Differenz von den Attributen Einnahmen und Ausgaben (Cleve und Lämmel 2024).

Ein weiteres Verfahren der Datentransformation ist die Aggregation. Laut Han et al. (2011) werden in der Aggregation Daten zusammengefasst und kompakter dargestellt. Das relevante an dieser Methode ist, dass die Zusammenfassung der Daten zu einem kleineren Datensatz

führt ohne wesentliche Informationen zu verlieren (Han et al. 2011). Cleve und Lämmel (2024) bezeichnen die Aggregation als Datenreduktionsverfahren und als Zusammenfassen von Fakten. Der Prozess der Aggregation umfasst die mathematischen Operationen, Summieren und Mittelwert Berechnung (Cleve und Lämmel 2024). Weiterhin behaupten Cleve und Lämmel (2024), dass die Aggregation auch als Dimensionsreduktion betrachtet werden könnte, da nicht nur Datensätze reduziert werden, sondern auch die Attribute selbst. Die Autoren betrachten die Dimensionsreduktion als Verfahren, bei der irrelevante Attribute vernachlässigt und nur relevante Attribute von Bedeutung sind. Die Frage bei dieser Methode ist, welche Attribute als irrelevant eingestuft werden können (Cleve und Lämmel 2024). Cleve und Lämmel (2024) unterscheiden zwei Situationen für die Dimensionsreduktion:

1. Wenn die Ausprägungen eines Attributs identisch sind, kann das jeweilige Attribut gelöscht werden
2. Wenn die Werte eines Attributs sehr unterschiedlich sind und bei der Nutzung eines Entscheidungsbaums keinen Mehrwert erschaffen.

Laut Alasadi und Bhaya (2017) existiert auch noch die Data-cube-Aggregation, bei der eine Operation über mehrere Tabellen abläuft und nicht wie die normale Aggregation eine simple Zusammenfassung ist.

Eine weitere Kategorie, die von Cleve und Lämmel (2024) untersucht wird, ist die Glättung**.** In der Glättung werden die ursprünglichen numerischen Werte einer Datenmenge durch idealisierte Werte ersetzt (Cleve und Lämmel 2024). Mit der Datenglättung wird eine reduzierte Menge angezielt, die zu besseren Lösungen führen soll (Cleve und Lämmel 2024). Cleve und Lämmel (2024) unterscheiden zwischen mehreren Datenglättungsmethoden:

- Binning
- Clustering
- Kombinierte menschliche und maschinelle Kontrolle
- Regression.

Im Binning werden Werte, die im gleichen Kontext auftauchen, in diskrete Gruppen eingeteilt (Han et al. 2011). Dabei werden die Nachbarwerte miteinander verglichen und in eine bestimmte Anzahl von Gruppen aufgeteilt (Han et al. 2011). Zu dem Prozess des Binnings gehört die Definition von Bereichen bzw. Intervallen, die die Anzahl der Gruppen beeinflusst (Han et al. 2011). Aggarwal (2017) führt weiter aus, dass die Gruppen von gleicher Größe sein sollten und durch die Mittelwerte repräsentiert werden. Der Autor behauptet, dass der Median der Gruppe auch benutzt werden könnte, da dieser robuster gegenüber Ausreißer ist. Han et al. (2011) ergänzt, dass die Werte auch durch die Grenzen der Gruppe repräsentiert werden können und diese durch die nächste Entfernung definiert werden. Das Hauptproblem des Binning ist die Methodik, die zu einem Verlust von verfügbaren Datenpunkten führt (Aggarwal 2017). Der folgende Datensatz wurde von Han et al. (2011) übernommen und soll die Vorgehensweise des Binning darstellen:

- Sortierter Datensatz: 4, 8, 15, 21, 21, 24, 25, 28, 34

Wie im vorherigen Abschnitt besprochen, müssen diese Werte in Gruppen aufgeteilt werden, die folgenderweise aussehen (Han et al. 2011):

- Gruppe 1: 4, 8, 15
- Gruppe 2: 21, 21, 24
- Gruppe 3: 25, 28, 34

Somit ist die Anzahl an Werten pro Gruppe identisch und können jetzt durch die Mittelwerte repräsentiert werden (Han et al. 2011):

- Gruppe 1: 9, 9, 9
- Gruppe 2: 22, 22, 22
- Gruppe 3: 29, 29, 29

Wie oben zu erkennen ist, wurde das Binning erfolgreich durchgeführt und somit ist die Datenglättung erfolgt (Han et al. 2011).

Das Clustering wurde in Abschnitt 2.3 als Methode des Data Mining eingeführt und kann auch als Datentransformationsverfahren benutzt werden (Cleve und Lämmel 2024). Hauptsächlich liegt hierbei der Fokus auf der Erkennung der Ausreißer und der nachfolgenden Behandlung (Cleve und Lämmel 2024). Die Vorgehensweise ist ähnlich wie im Data Mining, nur wird hier die Transformation des gegebenen Datensatzes durch die Entfernung bzw. Behandlung der Ausreißer thematisiert (Cleve und Lämmel 2024).

Eine weitere Datenglättungsmethode ist die kombinierte menschliche und maschinelle Kontrolle, wobei der Mensch den gegebenen Datensatz mit befremdlichen Werten auf Ausreißer untersucht und durch Expertise entfernt (Cleve und Lämmel 2024).

Die letzte Datenglättungsmethode ist die Regression und ist bereits wie das Clustering auch als Data-Mining-Methode schon vorgestellt worden. Mithilfe einer mathematischen Funktion werden Daten beschrieben und durch die berechneten Werte ersetzt (Cleve und Lämmel 2024). Hierbei ist der Fokus auf der linearen Regression mit der Zusammenhänge zwischen den Daten und den zugehörigen Funktionswerten verdeutlicht werden (Cleve und Lämmel 2024).

Nachdem die verfahrensunabhängigen Transformationen nach Cleve und Lämmel (2024) aufgezeigt wurden, werden jetzt weitere Verfahren vorgestellt, die im Kontext der Datentransformation auftauchen. Cleve und Lämmel (2024) thematisieren die Diskretisierung und verbinden dieses Verfahren mit der Umwandlung von metrischen in nominale oder ordinale Daten. Laut Han et al. (2011) werden die numerischen Rohdaten durch Intervalle oder durch konzeptuellen Labels ersetzt. Der Autor führt weiter aus, dass die Labels in höhere Hierarchien eingeordnet werden können und somit die numerischen Attribute wiedergeben. Alasadi und Bhaya (2017) behaupten, dass die Diskretisierung für die automatische Generierung von Hierarchien genutzt werden könnte. Die Autoren thematisieren im Kontext der Diskretisierung auch das Binning und sehen dieses als Diskretisierungsverfahren, da in beiden Methoden numerische Werte in größere Gruppen aufgeteilt werden. Ergänzend zum Binning werden Histogramme, Entropie basierte und Clustering als Diskretisierungsmethoden vorgestellt (Alasadi und Bhaya 2017). Mit Histogrammen werden nicht Labels verwendet, sondern Intervalle (Alasadi und Bhaya

2017). Han et al. (2011) führt aus, dass die Werte in gleichgroße Intervalle überführt werden und somit jeder Abschnitt die gleiche Anzahl von Daten besitzt. Durch die Wiederholung des Algorithmus können mehrere Hierarchien entwickelt werden (Alasadi und Bhaya 2017).

In der Methode der Entropie basierten Diskretisierung werden Klasseninformationen genutzt, um Datenmengen zu reduzieren (Alasadi und Bhaya 2017). Dabei wird das numerische Attribut X bzgl. ihrer minimalen Entropie untersucht und an diesem Punkt getrennt (Alasadi und Bhaya 2017). Wie in der Methode davor, wird dieser Schritt öfter wiederholt und eine Hierarchie erzeugt (Alasadi und Bhaya 2017).

Die letzte Methode einer Diskretisierung nach Alasadi und Bhaya (2017) ist das Clustering. In dieser Technik wird das numerische Attribut in Gruppen oder Clustern aufgeteilt (Alasadi und Bhaya 2017). Bei dieser Methode kann die Top-Down-Aufteilung oder auch die Bottom-Up-Zusammensetzung verwendet werden und ist somit in beide Richtungen anwendbar (Alasadi und Bhaya 2017). In diesem Fall der Diskretisierung können nicht nur höhere Hierarchieebenen gebildet werden, sondern auch niedrigere (Alasadi und Bhaya 2017). Die Cluster sind nicht wie im vorherigen Abschnitt, da sie wie Entscheidungsbäume aufgebaut werden und Knoten haben (Alasadi und Bhaya 2017). Ergänzend dazu werden diese überwacht, da sie Klassenbezeichnungen nutzen (Han et al. 2011).

Ein weiteres Verfahren zur Datentransformation ist die Normalisierung und Skalierung, die als Synonyme verwendet werden (Cleve und Lämmel 2024). Laut Cleve und Lämmel (2024) werden in diesen Verfahren Werte in einen bestimmten Wertebereich skaliert. Ergänzend zu den Begrifflichkeiten von Cleve und Lämmel (2024), wird vom Autor Runkler (2015) auch der Begriff Standardisierung benutzt. Allgemein werden bei der Normalisierung Werte in einen bestimmten Bereich angepasst (Alasadi und Bhaya 2017). Es existieren nach Cleve und Lämmel (2024) einige Verfahren und lauten:

- Min-Max-Normalisierung: Werte werden in einen Zahlenbereich zwischen 0 und 1 transformiert, wobei 0 den kleinsten Wert und 1 den höchsten Wert skaliert
- Z-Wert-Normalisierung: Die Differenz zwischen dem Wert und dem Mittelwert wird durch die Standardabweichung dividiert und führt zum normierten Wert
- Dezimalskalierung: Lineare Skalierung durch die Verschiebung des Kommas
- Logarithmische Skalierung: Logarithmus des Wertes bzgl. einer passenden Basis

Laut Cleve und Lämmel (2024) werden also gegebene Werte der Attribute in eine stetige, numerische Skala transformiert. Die Autoren Aggarwal (2017) und Runkler (2015) benutzen für die Z-Wert-Normalisierung den Begriff Standardisierung und stellen diese mit folgender Formel vor:

$$y_k^{(i)} = \frac{x_k^{(i)} - \mu^i}{\sigma^i}$$

Formel 7: Z-Wert Normalisierung

In diesem Kontext steht das μ für den Mittelwert und σ für die Standardabweichung (Aggarwal 2017). Der Wert x_k wird in diesem Fall normalisiert und wird dann durch y_k repräsentiert (Aggarwal 2017). Aggarwal (2017) erklärt dass die Mehrheit der Werte im Bereich zwischen -3 und 3 liegen werden. Außerdem greifen Aggarwal (2017) und Cleve und Lämmel (2024) auf die oben genannte Min-Max-Normalisierung zu und stellen diese mit der folgenden Formel vor:

$$x_{neu} = \frac{x - \min(x_i)}{\max(x_i) - \min(x_i)}$$

Formel 8: Min-Max-Normalisierung

Laut Cleve und Lämmel (2024) wird mit dieser Vorgehensweise auf das Intervall zwischen 0 und 1 normalisiert. Aggarwal (2017) führt aus, dass hierbei nur der minimale und maximale Wert des Attributs benötigt. Der Autor ergänzt, dass Probleme bei der Effektivität dieser Methode entstehen, wenn die maximalen und minimalen Werte so extrem sind, dass diese als Ausreißer bezeichnet werden können. Cleve und Lämmel (2024) fügen hinzu, dass der normierte Bereich verändert werden kann, indem diese Umwandlungsformel benutzt wird:

$$x_{neu=}a + (b - a) \times \frac{x - \min(x_i)}{\max(x_i) - \min(x_i)} \text{ mit dem normierten Intervall } [a, b]$$

Formel 9: Umwandlung der Normalisierung

Um die Normalisierung genauer zu erklären wird ein Beispiel von Han et al. (2011) vorgestellt. Han et al. (2011) stellt das Attribute Einkommen vor mit folgenden Informationen:

- Maximales Einkommen: 98.000$
- Minimales Einkommen: 12.000$
- Zu transformierender Wert: 73.600$

Han et al. (2011) stellen fest, dass der Wert zwischen 0 und 1 normiert werden soll und benutzten dafür die Formel (8), wodurch als Ergebnis 0,716 resultiert. Laut Alasadi und Bhaya (2017) ist die Normalisierung ein relevanter Prozess für verschiedene Data-Mining-Prozesse wie Klassifikation, künstliche neuronale Netze und Clustering. Ein Grund für die Relevanz ist, dass die Trainingsprozesse durch die neuen Wertebereiche verschnellert werden können (Alasadi und Bhaya 2017).

Runkler (2015) führt aus, dass die Auswahl der verschiedenen Verfahren abhängig von der gegebenen Verteilung der Daten ist. Außerdem ergänzt der Autor, dass bei der Auswahl der geeigneten Transformation auch der beobachtete und gewünschte Wertebereich von Bedeutung sind.

2.4.2 Entscheidungsunterstützung

Nachdem im letzten Unterabschnitt verschiedene Datentransformationsverfahren vorgestellt wurden, werden jetzt Entscheidungsunterstützungen zum Thema gebracht. Im Kontext dieser Arbeit soll eine Entscheidungsunterstützung zur Auswahl von Datentransformationsverfahren unter Berücksichtigung von den Bedingungen des Data Mining und der Automobilproduktion entwickelt werden. Dabei ist es relevant, wieso eine Entscheidungsunterstützung notwendig

ist, und welche Bedingungen berücksichtigt werden sollten. Dafür werden jetzt allgemeine Informationen zur Entscheidungsunterstützung aus der Literatur vorgestellt. Alpar et al. (2019) legen einen hohen Wert auf Informationen und auf den Prozess bevor eine Entscheidung getroffen werden kann. Informationen werden nach den Autoren als Rohstoff für den Entscheidungsprozess bezeichnet. Eine Situation zur Entscheidung entsteht, wenn zwischen mehreren Optionen gewählt werden muss (Pfister et al. 2017). Pfister et al. (2017) betonen, dass bei der Entscheidung nicht nur objektive Fakten in Betracht gezogen werden, sondern Emotionen und subjektive Erfahrungen die Wahl eines Menschen beeinflussen. Die Autoren ergänzen, dass es positiv wahrgenommen wird, wenn mehrere Optionen verfügbar sind, aber die Entscheidungsfindung bei unklaren Situationen zu negativen Emotionen führen kann. Um die Situationen anhand der Informationen zu verarbeiten, sind die kognitiven Fähigkeiten des Individuums von Bedeutung (Landes und Steiner 2013). Der Ausgangspunkt, um die Informationen zu verarbeiten, ist die Wahrnehmung (Landes und Steiner 2013). Mit der Wahrnehmung des Menschen werden Reize direkt und bestimmte Informationen unterbewusst verarbeitet (Landes und Steiner 2013). Mit der Wahrnehmung folgt auch das Wissen und Lernen des Menschen (Landes und Steiner 2013). Alle Reize, die wahrgenommen werden, werden durch den Menschen verarbeitet und fördern den Wissensstand und den Lernprozess (Landes und Steiner 2013). Diese kognitiven Fähigkeiten sind in den Vorphasen der Entscheidungsunterstützung essentiell (Alpar et al. 2019). Der Anwender bzw. der Mensch, der die Entscheidung trifft, muss anhand seines Vorwissens und seiner Gedanken entscheiden, welche Option passend für den Fall ist (Alpar et al. 2019).

Ein wichtiger Faktor bei der Entscheidungsunterstützung ist die Risikoeinschätzung, die die Persönlichkeit des Anwenders ausdrückt (Alpar et al. 2019). Die Risikoeinschätzung hängt vom Nutzenerwartungswert ab, der dem Anwender den Ertrag der Entscheidung wiedergibt (Alpar et al. 2019). Außerdem wirkt als weiterer Faktor die Sicherheitsäquivalenz, der die Sicherheit bzw. die Akzeptanz bzgl. eines Wertes für den Anwender wiederspiegelt (Alpar et al. 2019). Mit diesen beiden Faktoren wird dem jeweiligen Anwender eine Risikobereitschaft zugeordnet, die die Entscheidungsfindung beeinflusst (Alpar et al. 2019).

Durch die Definition von Zielen werden die Entscheidungsmöglichkeiten beschränkt (Pfister et al. 2017). Diese Ziele werden durch die kognitiven Fähigkeiten konstruiert (Pfister et al. 2017). Aufgrund von diesen Zielen werden also Entscheidungen getroffen, die dann vom Anwender begründet werden (Pfister et al. 2017). Mit dem Wissen der Ziele und der Gründe eines Anwenders lassen sich ihre Entscheidungen prognostizieren.

Alpar et al. (2019) führen nach der Erklärung von den Problemen der Entscheidungen, die Entscheidungsunterstützungssysteme (EUS) ein. Diese EUS ermöglichen den Anwendern eine Unterstützung bei Entscheidungsprozessen und dienen als Rationalitätssicherung (Alpar et al. 2019). Wie im Prozess der Entscheidung erklärt, besteht die erste Phase aus den vorhandenen Problemen, die zu optionalen Handlungen führen und wird als Planungsphase definiert (Alpar et al. 2019). Darauf folgt die Entscheidungsphase, bei der eine finale Entscheidung getroffen wird, nachdem in der letzten Phase mögliche Alternativen geprüft wurden (Alpar et al. 2019). Die letzte Phase wird als Durchsetzungsphase definiert und thematisiert die

Realisation bzw. Ausführung der Entscheidung mit der endgültigen Kontrolle (Alpar et al. 2019).

Die EUS können in den verschiedenen Phasen eingesetzt werden, um den Anwender bzgl. seiner Informationen zu helfen (Alpar et al. 2019). Alpar et al. (2019) unterscheiden zwischen zwei Kategorien von EUS und zwar zwischen daten- und methodenorientierten. Ersteres dient als Überwachungssystem und hilft bei der Aufbereitung, wobei die methodenorientierten Systeme problembelastete Entscheidungen mithilfe von Modellen bearbeiten (Alpar et al. 2019).

Alpar et al. (2019) behaupten, dass der Einsatz von EUS zu einer Effizienzsteigerung und kognitiven Entlastung führen können. Dennoch hinterfragen die Autoren den Gestaltungsprozess eines EUS, da festgelegt werden muss, ob deren Informationen aus subjektiven Einflüssen oder aus objektiven Informationen stammen wie aus dem Stand der Technik.

2.5 Automobilproduktion

Nachdem in den letzten Unterabschnitten Datentransformationsverfahren und allgemeines zu Entscheidungsunterstützungen besprochen wurden, wird die Datentransformation in den Domänenkontext der Automobilproduktion gestellt. Dafür werden allgemeine Informationen und der Nutzen von Daten in der Automobilproduktion zum Thema.

2.5.1 Allgemeines zur Automobilproduktion

Die Automobilindustrie ist ein essentieller Bestandteil der deutschen Wirtschaft (Fuchs und Westermeyer 2023). Laut Kropik (2021) ist die Automobilproduktion eine Sparte, die besonders durch ihre Dynamik ausgezeichnet ist. Der Autor führt weiter aus, dass die Dynamik durch den technologischen Fortschritt beeinflusst und somit Veränderungen entstehen. Diese Veränderungen beeinflussen nicht nur einzelne Unternehmen, sondern den ganzen Markt, da neue Trends entwickelt werden, die die Anzahl der Wettbewerber erhöht (Kropik 2021). Diese neuen Marktteilnehmer sorgen für Umstrukturierungen und fördern das Wachstum der Wirtschaft (Kropik 2021). Kropik (2021) führt aus, dass Fahrzeuge komplexe Produkte sind und für die Produktion mehrere Schritte erforderlich sind. In der folgenden Aufzählung werden die verschiedenen Fertigungstechnologien nach Kropik (2021) dargestellt:

- Im Presswerk werden Blechteile hergestellt durch Umformung
- In der Carbonfaser-verstärkten Kunststoff (CFK) Fertigung werden Karosserieteile hergestellt
- Karosserierohbau
- Lackiererei
- In der Teilefertigung werden aus spanender Fertigung/Spritzguss, neuerdings auch 3D-Druck bestimmte Kleinteile produziert
- Motormontage
- Batteriemontage für Elektrofahrzeuge
- Vormontagen von Einbauteilen, wie Sitzen
- Fahrzeugendmontage

- Zertifizierung durch Prüfung und Qualitätssicherung
- Auslieferung an Vertrieb.

Jeder einzelne Fertigungsschritt ist unterschiedlich geplant, automatisiert und verbraucht Produktionsressourcen (Kropik 2021). Diese Fertigungsschritte werden im Rahmen der Digitalisierung immer weiter entwickelt (Beckschulte et al. 2023). Kurpiela und Teuteberg (2023) führen aus, dass jedes Unternehmen der Automobilindustrie neue und bestehende Fahrzeugprojekte bzw. das Fahrzeugportfolio definieren und bewerten. Durch diese Vorgehensweise werden interne Projekte mit dem Wettbewerb verglichen und auf neue Trends vorbereitet (Kurpiela und Teuteberg 2023). Das Produktportfolio wird immer weiter erweitert und individueller für die Gestaltung der Kunden (Beckschulte et al. 2023). Wildemann (2013) behauptet, dass erst die Neustrukturierung der Produktionsbereiche zu Verbesserungen führen würden. Der Autor führt weiter aus, dass die Gegebenheiten in der Produktion stets hinterfragt werden und als Ausgangsbasis für weiter Verbesserungen sein sollen.

Mit der Digitalisierung findet also eine generelle Förderung der Automobilproduktion statt (Huber 2024). Kropik (2021) erklärt in diesem Kontext, dass die Komplexität der Fahrzeuge zu großen Informationsmengen führen. Es fällt auf, dass die Automobilproduktion sehr Datenabhängig ist (Huber 2024). Folgende Trends betreffen laut Fuchs und Westermeyer (2023) die Automobilproduktion:

- Neue Antriebssyteme (Elektromobilität)
- Additive Verfahren zur Fertigung
- Virtuelle Datennetze
- Autonomes Fahren.

Kropik (2021) verdeutlicht, dass bei allen neuen Trends Daten von großer Bedeutung sind, da mit den Daten eine Effizienzsteigerung erzielt werden soll. Weiterhin behauptet Kropik (2021), dass die vermehrte Nutzung der Digitalisierung den vorhandenen CO2-Ausstoß weiter erhöht, aber natürlich nicht der Hauptgrund dafür ist. Die Umweltbelastung ist in diesem Fall eines der Herausforderungen im digitalen Zeitalter der Automobilproduktion (Kropik 2021). Die größte Herausforderung durch die Digitalisierung ist aber die Datenmenge und deren ineffiziente Anwendung (Kropik 2021). Nachdem die Fertigungstechnologien und die Trends der Automobilindustrie thematisiert wurden, werden im nächsten Unterabschnitt die Relevanz von Daten und ihr Auftreten in der Automobilproduktion vorgestellt.

2.5.2 Daten in der Automobilproduktion

Mit der Digitalisierung verändern und entwickeln sich die Fertigungstechnologien weiter. Hierbei sind Daten essenzielle Bestandteile der Industrie. In diesem Unterabschnitt werden Daten in der Automobilproduktion genauer untersucht. Laut Alvarez-Coello et al. (2021) steigt das verfügbare Datenvolumen in der Automobilproduktion weiter an und somit auch die Rechenleistung sowie die Speicherkapazität. Der Autor führt weiter, dass Daten von Analysten als die wertvollste Ressource der Neuzeit betitelt werden, da diese digitale Innovationen und die Wirtschaft vorantreiben. Das Potenzial der Nutzung der Daten ist aber noch nicht völlig ausgeschöpft, da sie durch moderne Datenarchitektur erfolgen soll (Alvarez-Coello et al. 2021).

Fahrzeuge haben vor der Einführung des IoT ihre Daten lokal auf der Motorsteuerungseinheit gesammelt und verarbeitet (Giobergia et al. 2018). Mit diesen Daten wurden routinemäßige Vorgänge zur Verwaltung und Überwachung der Fahrzeugfunktionalitäten überprüft (Giobergia et al. 2018). Durch die begrenzten Kapazitäten dieser Motorsteuerungseinheiten wurden nicht alle vergangene Daten gesammelt (Giobergia et al. 2018). Mit den vernetzten Systemen der Neuzeit können jedoch mehr Daten gespeichert werden, da sie nicht nur lokal sind, sondern auch durch Cloud Möglichkeiten woanders gesichert wurden (Giobergia et al. 2018). Laut Schuh et al. (2019) ist die zunehmende Erfassung von Fahrzeugdaten von großer Bedeutung, da diese nicht nur Predictive Maintenance ermöglicht, sondern auch generell smarte Produkte. Diese Produkte sind nicht nur auf das Fahrzeug angepasst und verbessern das Fahrerlebnis des Kunden wie z. B. durch Unterhaltungselektronik (Schuh et al. 2019). Des Weiteren basiert die Vernetzung der Fahrzeuge nicht nur auf Datenaustausch und ermöglicht Updates (Schuh et al. 2019).

Die Fahrzeuge der digitalen Zeit besitzen neben allen mechanischen Komponenten auch komplexe Sensornetzwerke mit beeindruckenden Rechenressourcen (Alvarez-Coello et al. 2021). Dunckern (2017) gibt als Beispiel die Produktionsstandorte der BMW Group, die aus tausenden Rechnern und hunderttausenden sog. Produktionsendgeräten bestehen, die durch die Vernetzung den Produktionsprozess überwachen. Der Autor führt weiter aus, dass die Fahrzeuge mit den Connected-Drive-Tele-Services schon längst als digitale Produkte zählen. Nicht nur die Fahrzeuge sind digital, sondern auch die Maschinen und Anlagen in der Produktion (Dunckern 2017). Wie die Fahrzeuge besitzen die produzierenden Maschinen Software und Sensorik, die die Datenmenge erhöht und somit mehr Analysemöglichkeiten schafft (Dunckern 2017). Mithilfe der Sensorik können mehrere Fertigungsschritte automatisiert ablaufen (Ammal et al. 2021). Durch die Verbindung mit Künstlicher Intelligenz (KI) und Sensoren können die Maschinen ohne Menschen arbeiten (Ammal et al. 2021). Außerdem erkennt die KI mithilfe der Sensoren defekte und Abweichungen in der Produktion (Ammal et al. 2021). Laut Lee et al. (2023) sind Sensoren in der Fertigung und Prüfung von Fahrzeugen vorhanden und sind in digitale Produktionen essenziell. Durch die Verbindung der ganzen Produktionskette werden die Daten untereinander ausgetauscht (Lee et al. 2023).

Im folgenden Abschnitt werden die vorhandenen Daten in der Automobilproduktion genauer untersucht. Im Laufe der Zeit haben sich nicht nur die Daten entwickelt, sondern auch die Sensoren (Schütze und Helwig 2017). Die ersten Sensoren wurden als mechanische Messwertaufnehmer definiert, die sich dann weiterentwickelt haben und elektronisch wurden (Schütze und Helwig 2017). Schütze und Helwig (2017) definieren die Sensoren der heutigen Zeit als smart, da diese neben den verschiedenen Messmöglichkeiten auch noch selbstständiger geworden sind. Die Autoren führen weiter aus, dass die Sensoren Eigendiagnosen durchführen und bei Bedarf sich selbst konfigurieren und kalibrieren. Hering (2023) weist in diesem Kontext auf die Ausgangssignale hin, die vom Sensor ausgegeben werden und erklärt, dass das aufgenommene Signal erstmal umgewandelt wird. Der Autor führt weiter aus, dass durch Softwareprogramme eine Bearbeitung des Ausgangssignals erfolgt und dann zur Auswertung bereitsteht. Tränkler und Aronshtam (2014) erklären, dass je nach Messung eine an-

dere Verarbeitung erfolgt und somit auch andere Datentypen. Weiterhin thematisieren die Autoren, dass vor der Digitalisierung dieser Signale auch Qualitätsprüfende Schritte unternommen werden.

Neben den Sensordaten sind auch Stammdaten vorhanden. Gluchowski (2024) definiert Stammdaten als zustandsorientierte Daten, die über einen längeren Zeitraum zur Verfügung stehen. Laut Winkelhake (2024) sind Stammdaten ebenfalls relevant für innovative Lösungen. Für diese Stammdaten müssen Schnittstellen zwischen den verschiedenen Systemen geschaffen werden (Kropik 2021). Aufgrund von Freigabeprozessen ist die Übernahme von Stammdaten kein automatischer Prozess und benötigt eine Steuerung (Kropik 2021).

Zu den Stamm- und Sensordaten müssen auch Prozessdaten berücksichtigt werden (Possner et al. 2024). Durch die steigende Komplexität in den Produktionsprozessen entstehen immer wieder neue Daten, die als Prozessdaten bezeichnet werden können (Possner et al. 2024). Die Anwender bzw. Verantwortlichen dieser Prozesse sind aufgrund dieser Menge überlastet (Possner et al. 2024). Deswegen werden rechnergestützte und datengetriebene Methoden genutzt, die diese Prozessdaten auswerten (Possner et al. 2024).

Unabhängig von den oben genannten Datenarten unterscheiden sich je nach Prozess bzw. Fertigungsschritt die vorhandenen und neuerfassten Datentypen (Dietrich 2021). Dietrich (2021) thematisiert die Ablage von diesen Daten aus der Produktion und betrachtet Datenbanksysteme als geeignet. Der Autor führt weiter aus, dass diese Daten lokal gespeichert sein sollten und nur wenn sie aggregiert sind, an übergeordnete Systeme weitergegeben werden. Die Datenbanken für Sensoren sollen aus einfachen Tabellen bestehen, die mit einem Zeitstempel und dem erfassten Wert gespeichert werden (Dietrich 2021). Die erfassten Werte in den Datenbanken können verschiedene Datentypen beinhalten und hängen vom Sensor bzw. Prozess ab (Dietrich 2021).

Nachdem das Auftreten von Daten in der Automobilproduktion vorgestellt wurde, wird im nächsten Unterabschnitt die Datentransformation innerhalb der Domäne angeschnitten. Wie im letzten Abschnitt erwähnt, können sich die vorhandenen und erfassten Datentypen je nach Maschine bzw. Sensor unterscheiden. Wenn z. B. neue Prozessdaten mit Stammdaten verglichen werden sollten, könnten unterschiedliche Datentypen die Auswertung verfälschen. Um aber effizienter zu produzieren, müssen solche Qualitätsunterschiede bearbeitet werden. In diesem Kontext ist die Datentransformation und die verschiedenen Verfahren von Bedeutung, damit die Daten aus der Produktion umgewandelt werden können.

2.5.3 Datentransformationsverfahren in der Automobilproduktion

Das Ziel aller Wettbewerber ist das wirtschaftliche Wachstum. Dafür müssen die Daten des Unternehmens sinnvoll genutzt werden. Wie im Unterabschnitt davor genannt, können Daten die Effizienz der Produktion beeinflussen. Das Vorhandensein von Daten allein ist aber nicht ausreichend, denn die Daten müssen in einer Art und Weise auch Erkenntnisse bringen. Damit die Produzenten in der Automobilindustrie dies erreichen, müssen die Daten vorverarbeitet werden. In Abschnitt 2.4 wurde die Datentransformation eingeführt, die als Vorverarbeitungsschritt benutzt wird. Die Datentransformation erfolgt auch in der Automobilproduktion und wird

jetzt genauer thematisiert. Alvarez-Coello et al. (2021) erklären im Kontext der Daten in der Automobilindustrie, dass Daten in großer Menge vorhanden sind und aufgrund der Relevanz strukturiert analysiert werden müssen. Die Autoren definieren eine moderne Datenarchitektur in der Industrie, die Anhand von Regeln bzw. Standards organisiert werden sollte. Die vorhanden Daten soll unter Einhaltung dieser Regeln in einer Datenbank gespeichert werden (Alvarez-Coello et al. 2021). Um der Komplexität und der Menge dieser Daten entgegen zu wirken, ist eine Datenvorverarbeitung in der Produktion notwendig (Alvarez-Coello et al. 2021). Um das Nutzen dieser Daten zu erhöhen, muss das Verständnis verbessert werden. Aus diesem Grund ist die Datentransformation in den Datenbanken von der Automobilindustrie auch präsent (Alvarez-Coello et al. 2021).

Um die Nutzung der Datentransformation in der Produktion der Automobilindustrie darzustellen, werden verschiedene Beispiele präsentiert. Siek und Mukti (2020) führen ein Anwendungsbeispiel für die Produktionsdauer verschiedener Fertigungsschritte ein. Die Autoren betrachten dabei erworbenen Prozessdaten in der Produktion als zielführend für die Automatisierung. Dabei sollen die Prozessdaten in den Datenbanken klar definiert und mit Zeitstempel gekennzeichnet werden (Siek und Mukti 2020). Um die Analyse der Produktionsdauer durchzuführen, werden die grafischen Daten in eine Datenbank überführt und mithilfe von statistischen Methoden dargestellt (Siek und Mukti 2020). Die durchgeführten Transformationen bzw. Berechnungen fördern das Verständnis der Prozessdaten und vereinfachen dabei die Analyse.

Mane (2023) führt ein weiteres Anwendungsbeispiel aus der Produktion ein, bei der die Mitarbeiter Performance gemessen wird. Für die Messung werden verschiedene Verfahren des Data Mining mit vorheriger Datentransformation verwendet (Mane 2023). Viele verschiedene Attribute der Mitarbeiter werden ausgewertet, ohne auf die Skalenniveaus einzugehen (Mane 2023). Aus diesen Attributen resultiert eine einheitliche Tabelle, die die Bewertung der Mitarbeiter mithilfe einer Zahl von 1 (Schlecht) bis 5 (Exzellent) darstellt (Mane 2023). Unabhängig von dem Anwendungsbeispiel betont Mane (2023), dass statistische Verfahren für die Problemfindung und bessere Entscheidungen von Bedeutung sind.

Ein weiteres Beispiel ist von Kumari und Saini (2021), der auf die Daten von Sensoren eingeht. Neben Sensoren, die einfache physikalische Werte messen, existieren auch Sensoren die prüfen, ob Objekte vorhanden sind (Kumari und Saini 2021). Nach dem Sammeln der Daten werden verschiedene Datenvorverarbeitungsschritte durchgeführt, damit die Rohdaten nützlicher werden (Kumari und Saini 2021). Kumari und Saini (2021) behaupten, dass wenn Daten aus verschiedenen Quellen gewonnen werden, sie roh und nicht analysierbar sind. Die Autoren führen ihr Beispiel fort und wenden mehrere Datentransformationsverfahren hintereinander an. Außerdem können die Datenvorverarbeitungsschritte iterativ angewendet werden und somit auch mehrere Verfahren der Datentransformation (Kumari und Saini 2021).

Cupek et al. (2021) beschäftigen sich mit automatisch erfassten Daten von fahrerlosen Fahrzeugen. Diese setzt er mit Robotern gleich, die auch automatisiert in der Produktion arbeiten und durch die entsprechende Sensorik Daten erfassen (Cupek et al. 2021). Die Roboter bzw. Maschinen in der Produktion kommunizieren automatisch untereinander und tauschen somit Daten aus (Cupek et al. 2021). Diese Daten werden gesammelt und sind essentiell für den

Lernprozess der autonomen Systemen (Cupek et al. 2021). Als nächster Schritt erfolgt die Aggregation, in der die Daten inhaltlich zusammengefasst werden (Cupek et al. 2021). Cupek et al. (2021) führen weiter aus, dass verschiedene Aggregationsmethoden angewendet und durch ihre mathematische Funktion dargestellt werden können. Die Autoren erklären, dass danach weitere Datenvorverarbeitungsschritte erfolgen und am Ende im Data Mining Muster gesucht werden können.

Nachdem verschiedene Anwendungsbeispiele für Datentransformation aus der Automobilproduktion vorgestellt wurden, wird in diesem Abschnitt die Entscheidungsunterstützung in der Domäne im Fokus stehen. Das Ziel dieser Arbeit ist eine Entscheidungsunterstützung bzgl. Datentransformationsverfahren in der Automobilproduktion. Der Einsatz von Entscheidungsunterstützungen ist keine Seltenheit und wird im Folgenden für einen anderen Anwendungsfall vorgestellt. Ungermann et al. (2019) erklären, dass wissensbasierte EUS in der Produktion angewendet werden. Laut den Autoren ist die Komplexität in der Produktion höher den je aufgrund der Datenmenge. Anlagestörungen sind für z. B. Instandhalter problematischer geworden (Ungermann et al. 2019). Mithilfe von EUS sollen interaktive Lösungen für komplexe Entscheidungen gefunden werden (Ungermann et al. 2019). Ungermann et al. (2019) benutzen für ihr Beispiel ein Bayessches Netz zur Fehlerdiagnose, um Kausalitäten grafisch anzusehen. Mithilfe dieses Modells können die Anwender informativ und auch bei limitierten Daten Entscheidungen treffen(Ungermann et al. 2019).

Wie im oberen Abschnitt zu sehen war, sind bereits Entscheidungsunterstützungen in der Produktion vorhanden. In diesem Fall bezieht sich die Entscheidungsunterstützung auf Fehler in Anlagen. In Kapitel 2 wurden das Data Mining, verschiedene Datentransformationsverfahren, die Automobilproduktion und die Entscheidungsunterstützung vorgestellt. Auf diese Weise wurde die Grundlage für die systematische Literaturrecherche und für die Konzeptionierung der Entscheidungsunterstützung im Kontext dieser Arbeit geschaffen.

3 Systematische Literaturrecherche zu Datentransformations-verfahren

In diesem Kapitel wird die systematische Literaturrecherche zu den Datentransformationsverfahren durchgeführt, die in Abschnitt 2.4 vorgestellt wurden. Die systematische Literaturrecherche soll zu einer klaren und geordneten Struktur von Datentransformationsverfahren führen, die dabei einen transparenten Einblick in die Suche danach ermöglichen. Um die Literaturrecherche durchzuführen, wird die Methodik in Abschnitt 3.1 vorgestellt. Mithilfe der vorgestellten Methodik werden in Abschnitt 3.2 die Bedingungen der Literaturauswahl thematisiert. Dabei werden Filter benannt, die den Suchrahmen eingrenzen. In Abschnitt 3.3 wird dann die Literaturrecherche durchgeführt und veranschaulicht. Schließlich folgt in Abschnitt 3.4 die Kategorisierung der vorher extrahierten Datentransformationsverfahren.

3.1 Methodik der systematischen Literaturrecherche

Im Rahmen dieser Arbeit wird die Methodik der systematischen Literaturrecherche nach vom Brocke et al. (2009) durchgeführt. Die Autoren definieren ihren Prozess der Literaturrecherche über fünf Phasen. In der ersten Phase wird der Bereich der Literaturübersicht eingegrenzt (vom Brocke et al. 2009). Dabei wird festgelegt, welche Forschungsfragen und Ziele existieren (vom Brocke et al. 2009). Um das Vorgehen klar zu strukturieren, wird die Taxonomie nach Cooper (1988) eingeführt, die in Phase eins verwendet wird (vom Brocke et al. 2009). Die Taxonomie nach Cooper (1988) wird in Tabelle 4 dargestellt und sieht folgenderweise aus:

Tabelle 4: Taxonomie (in Anlehnung an Cooper (1988), S. 3-9)

Merkmal	Kategorien
(1) Fokus	Forschungsergebnisse & -methoden, Theorien, Anwendungen
(2) Ziele	Integration, Kritik, zentrale Aspekte
(3) Organisation	Historisch, konzeptionell, methodisch
(4) Perspektive	Neutrale Darstellung, vertretende Position
(5) Audienz	Spezialisierte Wissenschaftler, allgemeine Wissenschaftler, praxisorientierte, Öffentlichkeit
(6) Abdeckung	Vollständigkeit, vollständig selektiv, repräsentativ, zentral

Mithilfe der Taxonomie aus Tabelle 4 werden Forschungsfragen und somit der Schwerpunkt der Literaturrecherche definiert (Cooper 1988). Darauffolgend werden die Ziele heraus gearbeitet, wobei der Zweck in die Forschungslücke eingeordnet wird (Cooper 1988). Die Organisation definiert die Struktur innerhalb der Literaturrecherche. Hierbei wird nach Art (Historisch,

konzeptionell oder methodisch) die Übersicht angeordnet (Cooper 1988). Das Perspektivmerkmal umfasst die Sichtweise des Verfassers (Cooper 1988). Unter Audienz wird hinterfragt, an wen die Übersicht gerichtet ist (Cooper 1988). In der Abdeckung wird der Umfang der Literaturübersicht zum Thema gebracht und inwiefern die relevantesten Quellen berücksichtigt wurden (Cooper 1988). Nach der Taxonomie kann die zweite Phase der systematischen Literaturrecherche nach vom Brocke et al. (2009) bearbeitet werden. In der zweiten Phase wird der konzeptuelle Rahmen, also Schlüsselbegriffe und Datenbanken thematisiert (vom Brocke et al. 2009). Die dritte Phase ist die Durchführung der Literaturrecherche, bei der das Vorgehen dokumentiert und die Ergebnisse nachverfolgbar wiedergegeben werden (vom Brocke et al. 2009). In der vierten Phase wird die ausgewählte Literatur genauer analysiert, bzw. der Inhalt wird untersucht(vom Brocke et al. 2009). Dazu soll eine Konzeptmatrix erstellt werden, die die Literatur kategorisiert und bewertet (vom Brocke et al. 2009). In der letzten Phase werden die Erkenntnisse der Konzeptmatrix wieder in die Forschungsagenda eingeordnet und für die zukünftige Recherche bereitgestellt (vom Brocke et al. 2009).

3.2 Bedingungen der Literaturauswahl

Nachdem im letzten Abschnitt die allgemeine Vorgehensweise der systematischen Literaturrecherche nach vom Brocke et al. (2009) ausgearbeitet wurde, wird diese an der Thematik dieser Arbeit durchgeführt. Dazu werden die vorgestellten Phasen aus Abschnitt 3.1 im Kontext dieser Arbeit bearbeitet.

Phase 1

In der ersten Phase wird mithilfe der Taxonomie nach Cooper (1988) der Literaturbereich eingegrenzt. Im Folgenden werden die Merkmale aus der Taxonomie nacheinander bearbeitet.

- Fokus: In der systematischen Literaturrecherche werden Datentransformationsverfahren gesucht basierend auf den Verfahren aus Kapitel 2, die als Vorverarbeitung im Data Mining gelten. Im Optimalfall sollen die Verfahren aus dem Data Mining der Automobilproduktion stammen.
- Ziele: Es sollen Verfahren ermittelt werden, die ggf. die Basisverfahren aus Kapitel 2 erweitern und in der Automobilproduktion anwendbar sind. Neue Verfahren sind nicht ausgeschlossen.
- Organisation: Die Ergebnisse werden methodisch dargestellt, da die Verfahren im Fokus dieser Recherche sind.
- Perspektive: Innerhalb der Literaturrecherche erfolgt eine neutrale Stellung bzgl. allen Verfahren. Die Bewertung folgt nach der Literaturrecherche.
- Audienz: Die Ergebnisse der Literaturrecherche sind für die spezialisierten Wissenschaftler und auch für Data Scientisten in der Automobilproduktion.
- Abdeckung: Es erfolgt eine vollständig selektive Recherche, die aufgrund von Ausschlusskriterien erfolgt.

Nachdem die Taxonomie nach Cooper (1988) bearbeitet wurde, endet die erste Phase nach vom Brocke et al. (2009).

Phase 2

In der zweiten Phase werden jetzt die Schlüsselbegriffe und die Datenbank ausgewählt. Die Schlüsselbegriffe basieren in diesem Fall auf die verschiedenen Datentransformationsverfahren, die in Abschnitt 2.4 vorgestellt wurden. Die Schlüsselbegriffe lauten:

- Anpassungen/Codierung
    - One-Hot-Encoding/Binarization
- Kombinieren und Separieren von Attributen
    - Generalisierung
- Berechnung von abgeleiteten Werten
- Aggregation
- Glättung
    - Binning
    - Clustering
    - Kombinierte menschliche und maschinelle Kontrolle
    - Regression
- Diskretisierung
- Normalisierung/Skalierung/Standardisierung

Um zielführender zu sein, werden allgemeine Begriffe durch die spezialisierten Verfahren ersetzt. Dies betrifft die Begriffe Anpassungen/Codierungen, Berechnungen von abgeleiteten Werten und Kombinieren und Separieren von Attributen. Ein weiterer Grund die Gruppe Berechnung von abgeleiteten Werten zu entfernen, ist die Tatsache, dass keinen speziellen Datentransformationsverfahren im Kontext des Data Mining angewendet werden. In dieser Gruppe wurden simple mathematische Operationen durchgeführt, die auch ohne das Verständnis der Datentransformation angewendet werden können. Das spezialisierte Verfahren Kombinierte menschliche und maschinelle Kontrolle wird ebenfalls entfernt, da keine Verfahrensbeschreibung existiert. In diesem Fall wird nur das Zusammenwirken vom Menschen und von Maschinen angedeutet, die auch in den anderen Verfahren so gestaltet werden kann. Die Begriffe Diskretisierung und Normalisierung/Skalierung/Standardisierung sind auch wenn spezialisierte Verfahren existieren, alleinstehend als Datentransformationsverfahren zu erkennen. Folgende Schlüsselbegriffe der Datentransformationsverfahren resultieren nach der Eingrenzung:

- One-Hot-Encoding/Binarization
- Generalisierung
- Aggregation
- Binning
- Clustering
- Regression
- Diskretisierung
- Normalisierung/Skalierung/Standardisierung

Die Schlüsselbegriffe sind jetzt z. T. in deutscher und englischer Sprache, die aufgrund der Quellenvielfalt aus dem Stand der Technik so ausgewählt wurden. Um die systematische Literaturrecherche auszubreiten, werden alle Begriffe in die englische Sprache umgewandelt, weil die englische Sprache als internationale Sprache gilt und somit den größten Teil der Wissenschaft abdeckt. Ein weiterer Aspekt sind die Synonyme für die Verfahren. In der Bearbeitung aus dem Stand der Technik wurden für das OHE durch den Vergleich der Methodik das Synonym Binarization festgestellt. Weitere Synonyme wurden für die Normalisierung entdeckt und lauten Skalierung und Standardisierung. Innerhalb der systematischen Literaturrecherche werden diese Synonyme mithilfe des Booleschen Operator OR eingebunden. Die Schlüsselbegriffe werden nach den festgelegten Entscheidungen dargestellt:

- One-Hot-Encoding OR Binarization
- Generalization
- Aggregation
- Binning
- Clustering
- Regression
- Discretization
- Normalization OR Scaling OR Standardization

Diese Begriffe werden im ersten Abschnitt der Suchstrings eingebunden. Es werden also acht Suchstrings benutzt, die durch die Ergänzung von weiteren Schlüsselbegriffen entstehen. Der Boolesche Operator AND wird in diesen Fällen die Suchstrings erweitern und neben der Information des jeweiligen Datentransformationsverfahren die Suche durch weitere Schlüsselbegriffe eingrenzen. Im Kontext dieser Arbeit werden Datentransformationsverfahren in der Datenvorverarbeitung der Automobilproduktion gesucht. Das bedeutet, dass die Schlüsselbegriffe die Datenvorverarbeitung und auch ggf. die Automobilproduktion beinhalten sollten. Wie im Fall der Datentransformationsverfahren, müssen die Begriffe in die englische Sprache umgewandelt und dabei Synonyme berücksichtigt werden. Die Recherche aus dem Stand der Technik führte für die Datenvorverarbeitung zu zwei verschiedenen Begriffen. Aggarwal (2017) benutzt z. B. Data Preparation für die Datenvorverabeitung, wobei Fan et al. (2021) es als Data Preprocessing definiert. Beide Autoren beschreiben die Datenvorverarbeitungsschritte vor dem Data Mining und betrachten die Datentransformation als Inhalt. Folgende Schlüsselbegriffe resultieren aus dieser Erkenntnis:

- Data Preparation OR Data Preprocessing

Für Daten bzw. in diesem Fall Data existieren auch Synonyme, die aufgrund der Recherche aus Kapitel 2 ausgeschlossen werden. Alle Autoren bzw. Quellen, die in Kapitel 2 zitiert wurden, verwendeten den Begriff Data, wenn das Data Mining und die Datentransformation thematisiert wurden. Der Schlüsselbegriff Datentransformation wird nicht verwendet, da verschiedene Verfahren nicht unbedingt der Datentransformation zugeordnet wurden. Je nach Quelle gab es Unterscheidungen, ob Verfahren einzeln oder unter der Datentransformation vorgestellt wurden. In einigen Fällen wurden Verfahren unter anderen Datenvorverarbeitungsschritten vorgestellt. Deshalb werden die allgemeinen Begriffe der Datenvorverarbeitung benutzt, um

diese Verfahren nicht auszuschließen, wenn diese z. B. unter der Datenreduktion vorgestellt wurden.

Wie im oberen Abschnitt erklärt, wird in dieser Arbeit als Domäne die Automobilproduktion untersucht. Deshalb sollte es in Betracht gezogen werden, diese ebenfalls im Suchstring zu verwenden. Wie in den Fällen davor wird der Begriff in der englischen Sprache verfasst. Zum wiederholten mal wird untersucht, welche Synonyme für die Automotive Production verwendet werden können. In diesem Fall müssen Synonyme für beide Begriffe untersucht werden. Der Begriff Industrie, der im Laufe dieser Arbeit öfter mal als Synonym auftrat, würde in der systematischen Literaturrecherche die Suche aus dem Produktionskontext entfernen, da der Begriff weitaus mehr als die Produktion beinhaltet. Zielführender sind Begriffe wie Manufacturing oder Assembly, die man als Produktionsbegriffe bezeichnen kann. Für Automotive können Begriffe verwendet werden, die auf Fahrzeuge andeuten bzw. diese definieren. Folgende Begriffe resultieren:

- Automotive OR Automobile OR Vehicle OR Car
- Production OR Manufacturing OR Assembly

Diese Begriffe können in Kombination benutzt werden, um in der Suche die Automobilproduktion zu berücksichtigen.

Neben den Schlüsselbegriffen ist die Datenbank auch von Bedeutung. In der systematischen Literaturrecherche wird die Datenbank Scopus verwendet. Die Datenbank Scopus untersucht die Informationen aus naturwissenschaftlichen Quellen und Verlagen. Somit sind verschiedene Quellen in der Suche von Scopus schon beinhaltet. Außerdem wird ein Zeitraum für die Publikationen festgelegt, da in dieser Arbeit der Kontext des Data Mining und der Digitalisierung behandelt wird. Durch die stetige Entwicklung der Automobilproduktion, vor allem in den letzten Jahren, sollten die ermittelten Verfahren der jetzigen Zeit entsprechen. Unter Berücksichtigung der enormen Datenmengen und dem Trend von Big Data sollten Quellen der letzten 10 Jahre untersucht werden. Wichtige Informationen bzw. relevante Datentransformationsverfahren werden immer wieder von aktuelleren Quellen aufgegriffen und wiedergegeben. Bei Bedarf ist dann durch eine Rückwärtssuche der Literatur der Ursprung von den jeweiligen Datentransformationsverfahren auffindbar. Außerdem sollten Ausschlusskriterien erläutert werden. In diesem Fall können es inhaltliche Abweichungen sein, die durch die genaue Prüfung ermittelt werden. Des Weiteren müssen Fälle berücksichtigt werden, die durch die systematische Literaturrecherche erscheinen, aber aufgrund von Zugriffsrechten nicht geöffnet werden können. Da diese Arbeit an der Technischen Universität Dortmund durchgeführt wird, werden Dokumente untersucht, die durch die gegebenen Zugriffsrechte der Universität und unabhängig von Institutionen geöffnet werden können. Durch diese Bedingungen wurde eine Eingrenzung des Literaturbereichs unternommen, die die systematische Literaturrecherche vereinfachen soll.

3.3 Durchführung der Literaturrecherche

In diesem Abschnitt wird die systematische Literaturrecherche nach vom Brocke et al. (2009) durchgeführt. Im letzten Abschnitt wurden die ersten beiden Phasen bearbeitet und bilden die Grundlage der systematischen Literaturrecherche. In der dritten Phase findet die Durchführung der Literaturrecherche statt.

Phase 3

In dieser Phase werden die verschiedenen Suchstrings erschaffen und in der Datenbank Scopus mit dem bestimmten Zeitraum eingegeben. Der allgemeine Suchstring sieht folgenderweise aus:

("A" OR "B") AND ("Data Preparation" OR "Data Preprocessing") AND ("Automotive Production" OR "Automotive Manufacturing" OR "Automotive Assembly" OR "Automobile Production" OR "Automobile Manufacturing" OR "Automobile Assembly" OR "Vehicle Production" OR "Vehicle Manufacturing" OR "Vehicle Assembly" OR "Car Production" OR "Car Manufacturing" OR "Car Assembly")

Dieser Suchstring bezieht sich somit auf die Datentransformationsverfahren und beinhaltet dabei die Datenvorverarbeitung in der Automobilproduktion. In der ersten Klammer sind die Variablen A und B, die je nach dem gesuchten Datentransformationsverfahren auszuwählen sind. Je nach Anzahl der Synonyme entstehen oder fehlen Variablen in der ersten Klammer. An den Suchstring wird die oben genannte Bedingung des Zeitraums ergänzt:

AND PUBYEAR > 2013 AND PUBYEAR < 2025

Der letzte Teil des Strings sagt aus, dass die Veröffentlichungen erst ab 2014 zählen. Zudem werden zuerst nur Titel der Veröffentlichungen untersucht. Nachdem der erste Prototypsuchstring erstellt wurde, werden die Datentransformationsverfahren in die Variablen eingesetzt. Die erste Suche nur anhand der Titel führt zu keinen Ergebnissen, vor allem durch die Spezialisierung der Domäne und des Datenkontexts. Aufgrund dieser Tatsache wird der Suchstring auf Titel, Abstrakt und Schlüsselbegriffe erweitert, die durch die erweiterte Suche der Datenbank Scopus möglich ist. Mit der Einstellung resultierte diese Anzahl an Dokumenten:

- Generalization: 1
- Regression: 1

Die jeweiligen Strings führten nicht zum erwünschten Ergebnis. In diesem Fall wird der String nicht bzgl. der Suche nach Titel, Abstrakt und Schlüsselbegriff angepasst, sondern auf die Schlüsselbegriffe, die in Abschnitt 3.2 vorgestellt wurden. Die erzeugten Suchstrings sind sehr speziell bzgl. der Anwendungsdomäne gerichtet und sollten allgemeiner gestaltet werden. Dazu werden die Kombinationen der Begriffe, durch die einzelnen Begriffe Automotive, Automobile, Vehicle und Car erweitert, um mehrere Ergebnisse zu erzielen. Außerdem wird noch der Begriff der Datentransformation (Data Transformation) in den Suchstring ergänzt, um ebenfalls die Wahrscheinlichkeit von Ergebnissen zu erhöhen. Dennoch muss bei der inhaltlichen Prüfung untersucht werden, ob die hier definierten Datentransformationsverfahren unter

einem anderen Datenvorverarbeitungsschritt gelistet sind. Dadurch resultiert folgender Suchstring:

TITLE-ABS-KEY(("A" OR "B") AND ("Data Preparation" OR "Data Preprocessing" OR "Data Transformation") AND ("Automotive" OR "Automobile" OR "Car" OR "Vehicle" OR "Automotive Production" OR "Automotive Manufacturing" OR "Automotive Assembly" OR "Automobile Production" OR "Automobile Manufacturing" OR "Automobile Assembly" OR "Vehicle Production" OR "Vehicle Manufacturing" OR "Vehicle Assembly" OR "Car Production" OR "Car Manufacturing" OR "Car Assembly")) AND PUBYEAR > 2013 AND PUBYEAR < 2025

Der erste Abschnitt vor den Variablen ist durch die erweiterte Suche von Scopus entstanden und thematisiert die Suche nach Titel, Abstrakt und Schlüsselbegriffen. Des Weiteren wurden die Erweiterungen aus dem vorherigen Abschnitt übernommen. Dadurch resultieren für die verschiedenen Datentransformationsverfahren folgende Anzahl an Ergebnissen:

- One-Hot-Encoding/Binarization: 2
- Generalization: 26
- Aggregation: 10
- Binning: 0
- Clustering: 74
- Regression: 90
- Discretization: 3
- Normalization/Scaling/Standardization: 49

Die Anzahl der Ergebnisse ohne Inhaltsprüfung sind noch zu gering, vor allem für One-Hot-Encoding/Binarization, Aggregation, Binning und Discretization. Die Gründe für die geringe Anzahl sind erstens die Eingrenzung auf die Automobilproduktion und zweitens die Synonyme für die Verfahren. In Abschnitt 3.2 wurde die Entscheidung getroffen, dass spezielle Begriffe benutzt werden sollten, wie z. B. OHE statt Codierung. Diese Entscheidung wird verworfen, um die Domäne im Suchstring noch einzuhalten. Außerdem wird diese Entscheidung bei weiteren Datentransformationsverfahren durchgeführt und der spezielle Begriff mit der höheren Stufe erweitert. Im Fall des Binning wird also in den String das Synonym der Datenglättung mitberücksichtigt. Die Aggregation wurde im Unterabschnitt 2.4.1 auch als Dimensionsreduktion bezeichnet, die im weiteren Verlauf als Synonym in der Literaturrecherche verwendet wird. Die Diskretisierung wird mit dem Begriff der Categorization erweitert, da metrische bzw. numerische Daten in nominale verwandelt werden. So sehen die erweiterten Begriffe aus:

- One-Hot-Encoding/Binarization/Coding/Encoding
- Aggregation/Dimension Reduction
- Binning/Smoothing
- Discretization/Categorization

Mit diesen neuen Synonymen werden die aktuellen Strings ausgefüllt. Es resultieren folgende Ergebnisse:

- One-Hot-Encoding/Binarization/Coding/Encoding: 21
- Generalization: 26

- Aggregation/Dimension Reduction: 12
- Binning/Smoothing: 19
- Clustering: 74
- Regression: 90
- Discretization/Categorization: 6
- Normalization/Scaling/Standardization: 49

Aufgrund der geringen Anzahl von Ergebnissen wird der String komplett bearbeitet. Auch für die Normalisierung, das Clustering und die Regression sind die Anzahl an Ergebnissen nicht in Ordnung, wenn aus Erfahrungswerten berücksichtigt wird, wie viele der ersterzielten Ergebnisse nach genauer Prüfung noch übrigbleiben. Der Domänenabschnitt wird entfernt, da er zu spezifisch bei der Literaturrecherche ist. Stattdessen wird der Fokus auf den Kontext der Datenvorverarbeitung gesetzt. Es werden mehr Begriffe benutzt, die die Literaturrecherche auf die Datentransformationsverfahren spezifizieren. Das bedeutet, dass folgende Begriffe bedeutsam werden:

- Data Preparation, Data Preprocessing, Preprocessing Technique, Preprocessing Method, Preprocessing Algorithm. Data Transformation, Transformation Method, Transformation Technique, Transformation Algorithm, Feature Transformation, Feature Engineering

Außerdem werden die Datentransformationsverfahren nicht mehr einzeln gesucht, sondern in einem String. Das heißt, dass alle oben genannten Verfahren und Synonyme durch die Boolesche Operation OR verbunden werden. Diese Methodik soll dazu führen, dass die Systematische Literaturrecherche eine Menge an Publikationen findet, die auf das Vorkommen der Datentransformationsverfahren untersucht werden. Weiterhin werden Dokumente nach Titel, Abstract und Schlüsselwörtern untersucht. Außerdem folgt die Erkenntnis, dass die Begriffe Clustering und Regression in der Datenvorverarbeitung auftauchen, aber eher ein Teil des Data-Mining-Prozesses sind. Um Datentransformationsverfahren zu fokussieren, werden diese Begriffe in der Suche entfernt, aber in der Volltextprüfung wieder berücksichtigt, wenn diese in den Publikationen der Datentransformation auftauchen. Der neue String aus der erweiterten Suche von Scopus lautet:

TITLE-ABS-KEY ("Data Mining" AND "Data Transformation" AND ("One Hot Encoding" OR "Binarization" OR "Coding" OR "Encoding" OR "Generalization" OR "Aggregation" OR "Dimension Reduction" OR "Binning" OR "Smoothing" OR "Discretization" OR "Categorization" OR "Normalization" OR "Scaling" OR "Standardization") AND ("Data Preparation" OR "Data Preprocessing" OR "Preprocessing Technique" OR "Preprocessing Method" OR "Preprocessing Algorithm" OR "Data Transformation" OR "Transformation Method" OR "Transformation Technique" OR "Transformation Algorithm" OR "Feature Transformation" OR "Feature Engineering")) AND PUBYEAR > 2013 AND PUBYEAR < 2025 AND (LIMIT-TO (LANGUAGE , "English"))

Aus diesem String resultieren 56 Ergebnisse. Im Vergleich zum letzten String der mehr Ergebnisse lieferte, muss hier erwähnt werden, dass diese Ergebnisse eher im Kontext der Datenvorverarbeitung sind und somit zielführender. Des Weiteren waren bei der Einzelsuche der Begriffe die Wahrscheinlichkeit der Duplikate höher, weil Autoren bei der Thematisierung der

Datentransformation nicht nur von einem Verfahren berichten. Ergänzend zu den anderen Suchen, wurde in den Suchstring die englische Sprache als Ausschlusskriterium eingefügt. Die Dokumente werden mithilfe der Funktionalität der Datenbank Scopus exportiert und genauer untersucht. Mit dem Literaturverwaltungsprogramm JabRef wurde die exportierte Bib-Datei geöffnet. Durch die Funktionalität Find Duplicates unter Quality wurden Duplikate gesucht und eins entfernt. Danach wurden die Dokumente Alphabetisch nach dem Autor sortiert, um eine einfache Struktur zu erschaffen. Die Anzahl der Dokumente sank auf 55. Anhand ihres Titels und des Abstracts sollen diese bewertet und aussortiert werden. Das Ziel in diesem Schritt ist, dass aus den Publikationen eine Datentransformation in der Datenvorverarbeitung unabhängig vom Verfahren erfolgen und identifiziert werden muss.

Nachdem Titel und Abstract untersucht wurden, blieben 48 Publikationen interessant. Durch die strikte Erstellung des Suchstrings thematisieren die Dokumente größtenteils Anwendungen der Datentransformation aus einer spezifischen Branche.

Phase 4

Es wurde festgestellt, wie viele der Dokumente genauer untersucht werden müssen. Aus den Abstracts der Dokumente folgt, dass in ihrem Kontext eine Datentransformation mit einem bestimmten Verfahren durchgeführt wird. Dazu wird eine Konzeptmatrix erstellt, die auf einer Seite die Literatur darstellt und auf der anderen Seite das benutzte Verfahren. Ergänzend hierzu wird ein weiteres Feld mit Unbekanntem Verfahren bei den Datentransformationsverfahren hinzugefügt. Dieses Feld wird dann markiert, wenn eine Methode auftaucht, die nicht im Unterabschnitt 2.4.1 vorgestellt oder in der ersten Etappe der String Bildung entfernt wurde. Diese Vorgehensweise liefert verschiedene Datentransformationsverfahren, andererseits auch die Statistik der Nutzung. Ein weiterer Aspekt, der berücksichtigt wird, sind die Quellen der Dokumente. Wenn eine Quelle Datentransformationsverfahren angibt, wird überprüft, ob er den Ursprung des Verfahrens zitiert hat. Zur Vereinfachung der Übersicht werden die Autoren mit Jahr benannt und auf die Synonyme verzichtet. Wie bereits erwähnt, werden das Clustering und die Regression ebenfalls ausgewertet, auch wenn diese nicht im Suchstring benutzt wurden, da diese Methoden auch noch zum Data Mining zählen.

Im Anhang befinden sich 4 Tabellen, die für die Analyse der Literatur von Bedeutung sind. Die Tabelle A.1 stellt die Publikationen der systematischen Literaturrecherche dar. Die Tabelle A.2 bezieht sich auf die inhaltliche Analyse der Werke, dabei wird geprüft, welche Verfahren in den Publikationen angewendet werden. In der Tabelle A.3 werden die unbekannten Verfahren der Quellen benannt. Die Tabelle A.4 ist für die Rückwärtssuche, in der speziell geprüft wird, welche Quellen von den Publikationen benutzt wurden und welche Verfahren in diesen vorkommen. Die Quellen werden in der Rückwärtssuche nach Titel und Verweis in der Publikation ausgewählt mit dem Fokus auf die Methoden der Datentransformation und nicht auf den Anwendungsbereich.

Nachdem alle 48 Dokumente genauer untersucht wurden, resultierte, dass auf 3 Dokumente nicht zugegriffen werden konnte. Von den restlichen 45 Dokumenten wurden 13 aussortiert, da diese entweder keine Datentransformation beinhalten oder inhaltlich ein Datentransformation erwähnen, ohne es genauer zu thematisieren oder zu benennen. Die Problematik dabei

ist, dass durch die verschiedenen Synonyme Verfahren in verschiedene Datenvorverarbeitungsgruppen sortiert werden. Die resultierenden 32 Dokumente wurden in die erste Tabelle A.1 im Anhang vermerkt. Mit den 32 Dokumenten wurde dann eine Konzeptmatrix erstellt und mit den Verfahren aus dem Stand der Technik geprüft (s. Anhang A.2). Die resultierenden Verfahren sind aufgrund der Suchfilter sehr speziell und aktuell. An den meisten Titeln ist zu erkennen, dass die Datentransformation in bestimmten Kontexten der Wissenschaft angewendet wurden. Dies kann zur Schlussfolgerung führen, dass die Datentransformationsverfahren speziell für deren Anwendungsgebiet angepasst sind. Dennoch ist bei der Analyse aufgefallen, dass viele Titel sich auf die Grundlagen der Verfahren beziehen. Diese Verfahren wurden in ihrer simplen Form verwendet oder weiterentwickelt. Die weiterentwickelten Verfahren wurden nicht speziell auf den Anwendungsbereich limitiert. Das bedeutet, dass entwickelte Verfahren unter Berücksichtigung der eigenen Parameter in den anderen Anwendungsbereichen ebenfalls benutzt werden könnten. Bezüglich der Branche können die Datentransformationsverfahren also nicht kategorisiert werden. Folgende Abbildung stellt die Ergebnisse der Konzeptmatrix aus A.2 und aus der Tabelle A.4 in einem Diagramm dar:

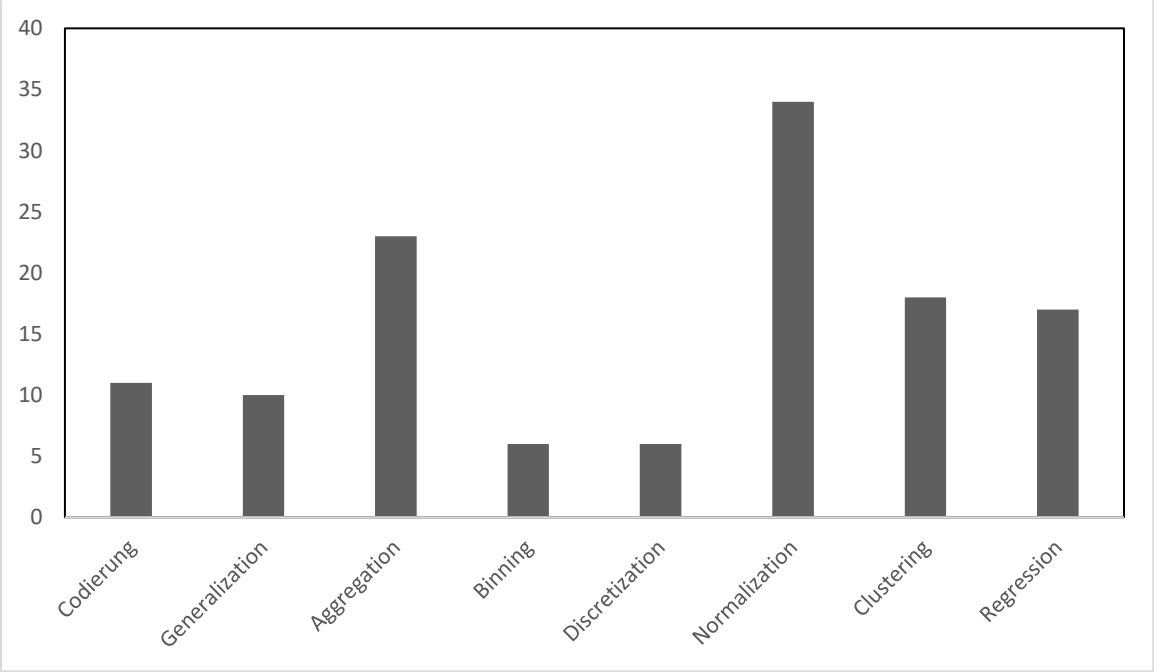


Abbildung 4: Statistik über die Nutzung der Datentransformationsverfahren pro Quelle

In der Abbildung wurden die Datentransformationsverfahren aus dem Stand der Technik bzgl. ihrer Nutzung in den Quellen der systematischen Literaturrecherche ausgewertet. Es ist zu erkennen, dass die Aggregation und die Normalization am häufigsten benutzt wurden. In der Auswertung wurden die spezifischen Formen der Datentransformationsverfahren auch berücksichtigt, indem diese direkt ihrem Hauptverfahren zugeordnet wurden. Im nächsten Abschnitt werden die Verfahren genauer untersucht und kategorisiert.

3.4 Kategorisierung der Datentransformationsverfahren

In diesem Abschnitt wird die vierte Phase der systematischen Literaturrecherche fortgeführt und die resultierenden Datentransformationsverfahren kategorisiert. Mit der Konzeptmatrix unter A.2 wurde ein Überblick der Verfahren aus der Literaturrecherche erschaffen. Interessanter sind die Erkenntnisse aus A.3, da Spezifizierungen der Datentransformationsverfahren zu sehen sind. In der Tabelle A.4 ist weitere Literatur, die aus den Quellen der systematischen Literaturrecherche sind. Diese sind größtenteils allgemeiner bzgl. der Datentransformation und haben keine Anwendungsdomäne. Diese Quellen wurden ebenfalls auf die vorhandenen Verfahren untersucht und dokumentiert. Mithilfe der beiden Tabellen A.3 und A.4 aus dem Anhang werden die Spezifizierungen der verschiedenen Datentransformationsverfahren dargestellt, dabei wird im weiteren Verlauf der Arbeit mit den englischen Begriffen gearbeitet:

- Aggregation:
    - Data-Cube-Aggregation
    - Mathematic Aggregation: (Summarize, Mean, Median and Variance)
    - Self- and Auxilary Aggregation (Text Aggregation)
    - Logical Aggregation
    - Dimension Reduction
- Normalization:
    - Min-Max-Normalization
    - Z-Scale-Normalization
    - Scaled-Clipping-Normalization
    - Column-Min-Max-Normalization
    - Global-Min-Max-Normalization
    - Column-Z-Score-Normalization
    - Global-Z-Score-Normalization
    - Quantile Normalization
    - Log Normalization
    - Euclidian Normalization
    - Blom Transformation
    - Euclidian-Distance-For-Similarity
    - Sigmoid Normalization
    - Gramian-Angular-Field-Encoding
    - Box-Cox-Transformation
    - Cube-Root-Normalization
    - Decimal-Scaling-Normalization
    - Normalization-Using-Norm
    - Square-Root-Normalization
    - Rank Transformation
    - Integer-Scaling-Normalization
- Mathematic Operations:
    - Logarithm
    - Square Root,
    - Squaring
    - Reciprocal
    - Axis/Spartial Transformation
    - Sine/Cosine Transformation
    - Linearization
    - Row Wise
    - Polynomial-/Non-Polynomial-Approximation-Of-Transformation
- Binning

    - Bin Means
    - Equal-Depth-Bins
    - Bin Boundaries
- Dizcretization
    - Binning
    - Histogram
    - Entropy-Based
    - Clustering

Mit den restlichen Verfahren Coding, Generalization, Clustering und Regression findet die Kategorisierung statt. Während der Analyse in der systematischen Literaturrecherche und der Ausarbeitung im Stand der Technik wurde festgestellt, dass die Datentransformation generell in zwei Gruppen aufgeteilt werden kann. Die erste Gruppe der Datentransformation beinhaltet die Verfahren, die eine komplette Veränderung des Datentyps ermöglichen (Externe Datentransformation). Das bedeutet, z.B. dass das Skalenniveau von nominal zu metrisch geändert wird. Beispielweise können mit dem Verfahren One-Hot-Encoding Begriffe einen Wert zugeordnet bekommen und somit das Skalenniveau der Daten verändern. Auf der anderen Seite existieren Datentransformationen, die innerhalb des Datentyps arbeiten (Interne Datentransformation). In diesen Verfahren werden Werte angepasst bzw. verändert. Diese Werte können mathematisch miteinander operieren und haben nicht das Ziel das Skalenniveau zu verändern. Des Weiteren findet eine Kategorisierung innerhalb dieser beiden Gruppen statt. Ein weiterer Aspekt, der beachtet werden muss, ist die Dimension des Datensatzes. Wie eben erwähnt sind Werte veränderbar. Dies geschieht durch unterschiedliche mathematische Operationen. In den Berechnungen der Mathematik sind Dimensionen vorhanden, die die Komplexität der Werte bzw. des Datensatzes beeinflussen. Eindimensionale Datensätze sind bzgl. des Aufwands anders zu kategorisieren als mehrdimensionale. Hierbei wird die Dimensionalität im mathematischen Sinne betrachtet und bezieht sich auf die Variablen im System. Aufgrund dessen werden die Datentransformationsverfahren in die Kategorien Limitierte Dimension und Mehrdimensionalität unterteilt. Die Verfahren, die für die Externe bzw. äußere Transformation zuständig sind, können auch eine Dimension haben, aber sind dann in der Dimension limitiert.

Die Tabelle 5 stellt die bisherige Kategorisierung der Datentransformationsverfahren dar.

Tabelle 5: Kategorisierung der Datentransformationsverfahren

Externe Transformation	**Interne Transformation**	
Gramian-Angular-Field-Encoding	**Limitierte Dimension**	**Mehrdimensionalität**
One-Hot-Encoding	Mathematic Aggr.	Data Cube Aggr.
Discretization	Self And Auxiliary Aggr.	Logical Aggr.
Discretization-Entropy-Based	Global-Norm.-Methods	Dimension Reduction
Discretization-By-Clustering	Column-Norm.Methods	Euclidian Norm.

Externe Transformation	**Interne Transformation**	
Discretization-By-Binning	**Limitierte Dimension**	**Mehrdimensionalität**
Discretization-By-Histogram	Quantile Norm.	Euclidian-Distance-For-Similarity-Norm.
Generalization	Log Norm.	Sigmoid-Norm.
	Blom Trans.	Normalization-Using - Norm
	Box-Cox-Trans.	Axis-/Spartial Trans.
	Cube-Root-Norm.	Sine-/Cosine Trans.
	Decimal-Scaling-Norm.	Linearisierung
	Square-Root-Norm.	Row Wise
	Rank Trans.	Polynomial-Approximation-Of-Trans.
	Integer-Scaling-Norm.	Non-Polynomial-Approximation-Of-Trans.
	Logarithm	Regression
	Square Root,	Clustering
	Squaring	
	Reciprocal	
	Bin Means	
	Bin Boundaries	
	Equal-Depth-Bins	

In der Tabelle 5 ist zu erkennen, dass die Datentransformationsverfahren der Encoding Varianten, die verschiedenen Arten der Discretization und die Generalization zu der Externen Transformation sortiert wurden. Mit diesen Verfahren können Skalenniveaus von Datensätzen so bearbeitet werden, sodass diese sich verändern. Bei der Gruppe Extern Transformation wird die innere Veränderung nicht ausgeschlossen. Eine Externe Transformation wird auch innerhalb des Datensatzes eine Veränderung bewirken. Als Beispiel werden im One-Hot-Encoding aus Begriffen Werte. Einerseits hat man die Transformation vom Nominalen ins Metrische, andererseits hat man die Veränderung des Inhalts, auch wenn dieser zugeordnet werden kann. Die Gruppe Interne Transformation beinhaltet jede mögliche Operation der Mathe-

matik. Außerdem wurden die Verfahren der Aggregation, Normalization, Clustering und Regression auch in die Interne Transformation sortiert. Diese Verfahren verändern den Datensatz im mathematischen Sinne und sind nicht für eine Veränderung der Skalenniveaus gedacht. Innerhalb der Verfahren der Aggregation, Normalization und der Mathematic Operations unterscheiden sich die Kategorien der Dimensionen. Durch die spezifizierte Form der Normalization wird ein eindimensionales Verfahren auch in der Mehrdimensionalität verwendet. Das gleiche gilt für die Aggregation. In der einfachsten Form der Aggregation, wie z. B. Summieren oder Mittelwertbildung, ist das Verfahren eindimensional. Aber durch die Erweiterung der Anwendung, wie z. B. das Summieren von Ergebnissen aus verschiedenen Tabellen, wird aus der Aggregation ein als mehrdimensionales Verfahren.

Die Kategorisierung der Datentransformationsverfahren resultierte aus der Konzeptmatrix und den weiteren Informationen aus den Tabellen der systematischen Literaturrecherche, die sich im Anhang A befinden. Die Kategorisierung spiegelt auch eine Bewertung der Verfahren. Die Verfahren werden nicht nur aufgrund externer und interner Transformation sortiert, sondern kriegen durch die Dimensionskategorie eine Komplexität zugeordnet. Diese Komplexität bezieht sich auf die Anwendung der Verfahren im mathematischen Sinne. Im Allgemeinen steigt die Komplexität, somit auch der Aufwand mit der Erhöhung der Dimension. Neben dem Aufwand, der durch die Dimension der Verfahren entsteht, muss auch individuell die Methodik bzw. das Vorgehen betrachtet werden. Falls Verfahren die gleiche Dimension haben, bedeutet das nicht, dass der Aufwand identisch ist. Mit dieser Kategorisierung endet die vierte Phase der systematischen Literaturrecherche.

Phase 5

Nachdem die Kategorisierung auch eine Bewertung der Verfahren widerspiegelt, werden jetzt die Ergebnisse in die Forschungsagenda eingeordnet. Der Fokus in dieser systematischen Literaturrecherche waren Datentransformationsverfahren, die auf den Schlüsselbegriffen aus Unterabschnitt 2.4.1 basieren. In der ersten Phase wurde es als Optimalfall definiert, wenn diese Verfahren direkt aus dem Data Mining der Automobilproduktion stammen. Darauf aufbauend wurden die ersten Prototypen der Suchstrings kreiert, die nicht direkt erfolgreich waren. Dennoch wurde das definierte Ziel erreicht, da die Ermittlung von Erweiterungen der Verfahren funktionierte, die in Automobilproduktionsdaten anwendbar sein sollten. Die Kategorisierung dieser Verfahren wird benutzt, um das Forschungsziel dieser Arbeit zu erreichen. Das Forschungsziel dieser Arbeit ist die Entwicklung der Entscheidungsunterstützung der verschiedenen Datentransformationsverfahren unter Berücksichtigung der Domäne. Dieses Ziel wird im nächsten Kapitel weiter erarbeitet.

4 Entwicklung einer Entscheidungsunterstützung zur Auswahl eines Datentransformationsverfahrens

Nachdem in Kapitel 3 die systematische Literaturrecherche erfolgte, wird in diesem Kapitel das Hauptziel dieser Arbeit thematisiert. Als Ziel in dieser Arbeit wurde die Entwicklung einer Entscheidungsunterstützung für die Auswahl der passenden Datentransformationsverfahren in der Automobilproduktion definiert. In Kapitel 2 wurden dafür grundlegende Informationen bzgl. Datentransformationsverfahren, Entscheidungsunterstützung und Automobilproduktion festgestellt. Die erweiterten Erkenntnisse aus der systematischen Literaturrecherche, das Wissen bzgl. Entscheidungsunterstützung und Automobilproduktion werden in diesem Kapitel zusammengeführt, um das Hauptziel zu erreichen. Vor der Konzipierung der Entscheidungsunterstützung werden dafür die Kriterien zur Anwendung der Datentransformationsverfahren bestimmt. Danach erfolgt eine Analyse der Anforderungen des Data Mining und der Randbedingungen der Automobilproduktion an die Datentransformationsverfahren. Mit diesen Informationen folgt in Abschnitt 4.4 die Entscheidungsunterstützung.

4.1 Kriterien zur Anwendung der Datentransformationsverfahren

In diesem Abschnitt werden die Kriterien festgelegt, die die Anwendung der Datentransformationsverfahren bestimmen. Zuletzt wurde in Abschnitt 3.4 mithilfe der Kategorisierung auch die Anwendung der Verfahren diskutiert. Ein Verfahren wird ausgewählt, wenn der Datensatz vorliegt. Das bedeutet, dass der Anwender sich bestimmte Merkmale des Datensatzes anschauen und basierend darauf sich für sein Verfahren entscheiden muss. Im Fall der besprochenen Thematik aus Abschnitt 3.4 muss das Verfahren an die Dimension des Datensatzes gerichtet sein und dem Anwender bewusst werden, wie der Zieldatensatz aussehen soll. Dies führt wieder zu der Kategorisierung. Einerseits kann das Skalenniveau verändert werden, andererseits nur das Innere des Datensatzes wie z. B. mit einer Min-Max-Normalisierung. Das Ziel und die Gegebenheit des Datensatzes sind die Hauptkriterien der Anwendung der Datentransformationsverfahren. Somit wird die Dimension als erstes Kriterium zur Anwendung der Datentransformationsverfahren definiert und bezieht sich auf die Dimension im mathematischen Sinne des vorliegenden Datensatzes. Das zweite Kriterium wird als Ziel definiert und beinhaltet die Vorstellung der finalen Form des Datensatzes.

Wie in Abschnitt 3.4 führen diese Aspekte unabhängig von der Dimension zum Aufwand. Der Aufwand eines Verfahrens muss stets beachtet werden. Der Aufwand kann sich in verschiedenen Formen bemerkbar machen. Einerseits durch lange Rechenzeiten oder auch durch Entscheidungen, bei dem der Anwender überlegen muss. Der Aufwand hängt also vom Verfahren und der Dimension des Datensatzes ab. Des Weiteren wird der Aufwand auch von der vorhandenen Datenmenge beeinflusst. Die Datenmenge in Kombination mit der vorhandenen Dimension kennzeichnen die Gegebenheiten, die den Aufwand ausmachen. Die Auswahl des Verfahrens wird anhand dieser Gegebenheiten entschieden und bringen einen eigenen Aufwand mit sich. Diese Aspekte machen das dritte Kriterium Aufwand aus.

Des Weiteren muss bei der Zieldefinierung festgelegt werden, wie Datenqualität definiert wird. Die Datentransformation kann den Inhalt beeinflussen und somit auch die Datenqualität. Im Unterabschnitt 2.1.2 wurden die Datenqualitätsdimensionen nach Strong et al. (1997) definiert. Die Datenqualitätsdimension beziehen sich im Allgemeinen nicht auf die Datentransformation. Dennoch ist es von Interesse, welche der Datenqualitätsdimensionen die Datentransformation bewerten könnten. Die Kategorie Zugänglichkeit und die vorhandenen Datenqualitätsdimensionen können komplett ausgeschlossen werden, da der Anwender als administrative Kraft stets Zugriff haben sollte. Aus der Kategorie Intrinsisch sind die Datenqualitätsdimensionen Glaubwürdigkeit und Reputation ebenfalls nicht geeignet als Kriterien. Der Grund dafür ist, dass die Herkunft der Daten aus der Automobilproduktion bekannt ist und der Anwender somit kein Misstrauen haben muss. Die Datenqualitätsdimensionen Genauigkeit und Objektivität werden genauer als mögliche Kriterien untersucht.

Die Dimension Genauigkeit könnte im Kontext der Datentransformation geprüft werden, wenn untersucht wird, inwiefern die Informationen nach der Transformation erhalten bleiben. Somit wird das vierte Kriterium als Informationsgehalt definiert. Dieses Kriterium bestätigt die Richtigkeit der Datentransformation, indem der Informationsgehalt geprüft wird. Der Informationsgehalt kann sich vor allem je nach Verfahren unterscheiden unter Berücksichtigung der Oberkategorien aus der Kategorisierung der Datentransformationsverfahren aus Abschnitt 3.4.

Die Beurteilung, ob eine Information nach der Transformation noch genau ist, kann je nach Anwender variieren. Also ist die Objektivität als generelles Kriterium ebenfalls ein Aspekt, welcher berücksichtigt werden muss. Die Auswahl der Methode und in der Methodik selbst müssen Entscheidungen getroffen werden, die entweder auf Intuition oder Erfahrung basieren. Mathematische Operationen können im Vergleich zur Generalization objektiv erfolgen. In der Generalization z. B. muss der Anwender eine Repräsentation wählen, wenn keine Vorgegeben ist. Diese Auswahl kann zu Missverständnissen führen, die vornherein abgeklärt werden müssen. Des Weiteren kann es auch bei vorgegebenen Ebenen zu Missverständnissen kommen, wenn die Anwender nicht untereinander absprechen, wie hoch Attribute generalisiert werden sollten. Beispielsweise kann das Attribut Straße in die Ebene Stadtteil oder auch Stadt eingeordnet werden. Die Einordung hängt in diesem Fall vom gewünschten Detaillierungsgrad des Anwenders ab. Im Unterabschnitt 2.4.1 wurde z. B. die Data-Cube-Aggregation eingeführt, die eine Aggregation von mehreren Tabellen thematisiert. In diesem Fall kann der Anwender entscheiden, wie die zusammengefasste Tabelle auszusehen hat. Somit entscheidet der Anwender auch wie kompakt er in der Methode zusammenfassen möchte. Somit wird das fünfte Kriterium als Objektivität definiert.

Bei der Betrachtung der weiteren Datenqualitätsdimensionen hat die Kategorie Repräsentativ im Kontext der Kriterien keinen Einfluss, da die Dimensionen Interpretierbarkeit, Verständlichkeit, Prägnanz und Konsistenz als vorhanden eingeschätzt werden. Des Weiteren existiert hier eine Überschneidung mit dem Kriterium Informationsgehalt, das die Aspekte dieser Datenqualitätsdimensionen berücksichtigt. Die Kategorie Kontextual mit den Datenqualitätsdimensionen Relevanz, Wert steigernd und Aktualität werden bei den Kriterien zu Anwendung nicht berücksichtigt. Dennoch wurden die Dimension Vollständigkeit und Datenmenge aus der Kategorie

Kontextual miteinbezogen. Die Dimension Vollständigkeit ist ebenfalls mit dem Kriterium Informationsgehalt abgedeckt und die Dimension Datenmenge wurde im Kriterium Aufwand berücksichtigt.

Basierend auf der Anwendung ist ein weiteres Kriterium, ob die Datentransformationsverfahren öfter an dem Datensatz und in Kombination mit einem anderen Verfahren benutzt werden können. In der Datenvorverarbeitung erfolgen viele Schritte, vor allem in der Datentransformation. Im Kriterium Kombination wird geprüft, ob das Verfahren mehrmals angewendet werden kann und welches Ergebnis daraus resultiert. Des Weiteren unterscheiden sich aufgrund der Resultate, ob unterschiedliche Verfahren nochmal an dem Datensatz benutzt werden können. Im Kriterium Kombination ist es also von Bedeutung, welches Ziel definiert wurde und welches Skalenniveau gegeben ist. Das Kriterium Kombination schließt die Nutzung vom gleichen Verfahren hintereinander nicht aus.

In der Datenvorverarbeitung existieren weitere Schritte außerhalb der Datentransformation. Diese Schritte könnten ebenfalls einen Einfluss auf die Anwendung der Datentransformationsverfahren haben. Viel mehr könnten weitere Datenvorverarbeitungsschritte als notwendig betrachtet werden. Im Data Mining und in der Datenvorverarbeitung wird die Datentransformation nie alleinstehend betrachtet. Wie in Abschnitt 2.4 erwähnt, kann vor oder nach der Datentransformation eine Datenreinigung oder auch eine Ausreißer Erkennung erfolgen. Diese Prozesse werden mit dem Kriterium Vorverarbeitung definiert.

In der folgenden Tabelle 6 werden die Kriterien nochmal aufgeführt und mit einem Bewertungsschema ergänzt.

Tabelle 6: Kriterien zur Anwendung der Datentransformationsverfahren

Kriterium	Bewertungsschema
Dimension	Ein positiver diskreter Wert.
Ziel	Ein bestimmtes Skalenniveau oder Wertebereich Transformation.
Aufwand	(Niedrig = 1; Mittel = 2; Hoch = 3).
Informationsgehalt	(Fehlgeschlagen; Erfolgreich).
Objektivität	(Niedrig = 1; Mittel = 2; Hoch = 3).
Kombination	Nennen der Verfahren und Anzahl der Wiederholungen (positiver diskreter Wert), falls gleiches Verfahren mehrmals verwendet werden.
Vorverarbeitung	Angabe der benötigten Vorverarbeitungsschritten.

Die Kriterien beeinflussen sich z. T. gegenseitig und wirken je nach Anwendung unterschiedlich aufeinander ein. Außerdem haben einige Kriterien unterschiedliche Bewertungsschemas zur Einschätzung, die jetzt genauer thematisiert werden.

Das Kriterium Dimension ist ein positiver diskreter Wert, der durch die Gegebenheit des Datensatzes erkennbar ist. In Abschnitt 3.4 wurden in der Kategorisierung die Verfahren bzgl. ihrer Dimensionen gruppiert. In der Anwendung eines Datentransformationsverfahren muss also die Dimension des Datensatzes bearbeitbar sein. Dieses Kriterium ist im mathematischen Sinne definiert und muss im Fall von nominalen Daten anders betrachtet werden. Im Unterabschnitt 2.1.1 wurden relationale Datenbanken thematisiert, die aus verschiedenen Tabellen bestehen. Nominale Daten können in diesen Tabellen auftreten und mehrere Felder haben. Wenn Felder innerhalb einer Tabelle einer bestimmten Variable zugeordnet werden können, sind diese Felder auch als Dimension zu betrachten. In einer Tabelle für die Daten von Fahrzeugen könnten z. B. Hersteller, Fahrzeugmodell, Seriennummer und Baujahr als Dimensionen der Variable Fahrzeug gelten. In diesem Fall wären die einzelnen Dimensionen einem Primärschlüssel zugeordnet, um eine Eindeutigkeit der Fahrzeuge zu gewährleisten. Das genannte Bewertungsschema für das Kriterium Dimension kann bei nominalen Daten somit durch die Anzahl der Spalten in der Tabelle angegeben werden und wären somit auch ein positiver diskreter Wert.

Das Kriterium Ziel basiert auf der Kategorisierung aus dem Abschnitt 3.4 und soll angeben, welches Skalenniveau aus dem Verfahren resultieren soll. Wenn das Skalenniveau nicht geändert werden soll, wird mit dem Ziel eine Transformation des Wertebereichs definiert. Anhand des angegebenen Ziels kann der Anwender untersuchen, welche Datentransformationsverfahren das definierte Ziel ermöglichen können. Mit der Hilfe der anderen Kriterien können weitere Datentransformationsverfahren in Betracht gezogen bzw. ausgeschlossen werden.

Das Kriterium Aufwand wird mit den Stufen Niedrig, Mittel und Hoch bewertet. Dabei ist zu beachten, dass verschiedene Kriterien bzw. Gegebenheiten den Aufwand beeinflussen. Das Kriterium Dimension und Ziel sind die Haupteinflüsse und geben die ersten Ansätze bzgl. des Aufwands an. Der Einfluss des Kriteriums Dimension ist eine generelle Betrachtung aus der Mathematik. Je höher die Dimension des Datensatzes ist, desto höher ist auch der Aufwand. Der Einfluss des Kriteriums Ziel ist unterschiedlich zu messen. Erstens muss geprüft werden, ob eine Skalenniveauveränderung oder eine Wertebereich Transformation stattfindet. Bei der Skalenniveauveränderung basiert der Aufwand nicht unbedingt auf mathematische Operationen.

Das Kriterium Informationsgehalt hat als Bewertungsschema die Angaben Erfolgreich und Fehlgeschlagen. Anhand der absolvierten Datentransformation erfolgt eine Bewertung der verwandelten Informationen. Der Anwender muss mithilfe des Bewertungsschemas angeben, inwiefern die Datentransformation abgelaufen ist. Die Bewertung Erfolgreich wird gegeben, wenn die vorhandenen Informationen nach der Datentransformation den Mehrwert für die Weiterverarbeitung beibehalten. Die Bewertung Fehlgeschlagen ist für die Fälle, wenn die Datentransformation nicht die gewünschten Ergebnisse erbracht hat und somit umgeplant werden sollte. Dennoch muss beachtet werden, dass die Einschätzung des Informationsgehalts vom Anwender abhängig ist.

Des Weiteren muss auch das Kriterium Objektivität ein zu berücksichtigender Faktor sein, der eine Bewertung erfordert. Die Bewertung basiert auf folgende Einflüsse. Wie oben bei der

Vorstellung der Kriterien erwähnt, gibt es Unterscheidungen zwischen den Verfahren. Der Anwender muss nicht nur zwischen den verschiedenen Datentransformationsverfahren eine Entscheidung treffen, sondern auch innerhalb dieser Verfahren. Die Objektivität der Anwender wird mit den Stufen Niedrig, Mittel und Hoch bewertet. Wobei Verfahren die keine Entscheidungsfreiheit zu lassen auch mit der Stufe Niedrig gekennzeichnet werden. Je anspruchsvoller bzw. größer die Datenmenge ist, desto mehr Aufwand wird benötigt. Im Fall der Generalization bedeutet das, dass mehrere Ebenen vorhanden sein können und somit auch mehr Entscheidungen im Verfahren notwendig sind. Dies kann dazu führen, dass nicht nur das Verfahren allein bzgl. der Objektivität bewertet werden muss, sondern es auch abhängig vom gegebenen Datensatz sein kann. Somit kann z. B. die Generalization je nach Anwendungsfall als Bewertung der Objektivität die Stufe Mittel oder auch die Stufe Hoch bekommen.

Das Kriterium Kombination wird mit der Nennung der nötigen Schritte gekennzeichnet. Die Bewertung erfolgt somit über die benutzten Datentransformationsverfahren und deren Anzahl der Anwendung. Ein Datentransformationsverfahren kann auch mehrmals angewendet werden und wird dann durch einen positiven diskreten Wert wiedergegeben.

Im Kriterium Vorverarbeitung findet die Nennung der Datenvorverarbeitungsschritte statt bevor es zur Anwendung des Datentransformationsverfahrens kommt. In Abschnitt 2.4 wurden weitere Datenvorverarbeitungsschritte genannt, die entweder vor oder nach der Datentransformation erfolgen können. Ergänzend zu den genannten Schritten existieren noch weitere Datenvorverarbeitungsschritte, die nicht weiter erläutert wurden. Im Falle der Bewertung wird der bekannte Datenvorverarbeitungsschritt genannt und anhand der Notwendigkeit durchgeführt.

Mit diesen Kriterien können die Datentransformationsverfahren genauer bzgl. des Anwendungsfalls ausgewählt werden. Die Kriterien aus diesem Abschnitt und die Kategorisierung der Datentransformationsverfahren aus dem Abschnitt 3.4 stellen bis jetzt die Basis der Entscheidungsunterstützung dar und thematisieren die Datentransformationsverfahren. Da die Entscheidungsunterstützung nicht nur von den Datentransformationsverfahren abhängig ist, werden im nächsten Schritt die Faktoren aus dem Data Mining vorgestellt.

4.2 Anforderungen des Data Mining an Datentransformationsverfahren

Nachdem im letzten Abschnitt Kriterien alleinstehend zur Anwendung der Datentransformationsverfahren entwickelt wurden, werden jetzt Anforderungen aus dem Data Mining vorgestellt. Die Datentransformation ist ein Vorverarbeitungsschritt und erfolgt deshalb vor dem Analyseprozess des Data Mining. Das bedeutet, dass die Datentransformation so ausgewählt werden, sollte, sodass der Analyseprozess erfolgen kann. In Abschnitt 2.2 wurden Herausforderungen des Data Mining erläutert, die auch Anforderungen an die Datentransformationsverfahren stellen können. Zum Analyseprozess gehören die verschiedenen Data-Mining-Methoden. Deswegen ist es auch von Bedeutung zu wissen, wie die Datensätze vor dem Anwenden der Data-Mining-Methoden vorbereitet sein sollten. Des Weiteren werden die Anforderungen der Data-Mining-Methoden vorgestellt.

Die Data-Mining-Methode beeinflusst das angezielte Skalenniveau der Datentransformation. Wie in Abschnitt 2.3 zu erkennen ist, ist nicht jedes Skalenniveau mit jedem Data-Mining-Prozess kompatibel, bzw. führt zu einer guten Analyse. Im Fall der Klassifikation oder der Assoziationsanalyse können z. B. nominale Daten vorhanden sein, die bearbeitet werden sollen. Weitere Daten, die zur Analyse dazukommen, müssen für die Vergleichbarkeit der Merkmale ebenfalls in das Skalenniveau transformiert werden, um den Informationsgewinn zu verstehen und nach der Analyse zu interpretieren. Die Anforderung wird als Skalenniveauanpassung definiert

Der Informationsgewinn im Data Mining kann nur erfolgen, wenn in der Vorverarbeitung relevante Information erhalten bleibt. Dies ist eine weitere Anforderung des Data Mining an die Datentransformationsverfahren, welches durch das Kriterium im vorherigen Abschnitt schon als wesentlich eingestuft wurde. Die Methoden aus dem Data Mining können neben der Anforderung eines bestimmten Skalenniveaus auch bestimmte Wertebereiche anfordern. Das bedeutet, dass auch innerhalb des Datentyps Anforderungen gestellt werden können. Während der Wertebereichveränderung müssen weitere Aspekte der Datentypen beachtet werden. In Abschnitt 2.1 wurden Standarddatentypen und ihre Bezeichnung in der Programmiersprache Java gegenübergestellt. Jeder Standarddatentyp hat einen gegeben Wertebereich, der durch die Datentransformation über- oder unterschritten werden kann. Das Data Mining und die Datentransformation sind Prozesse, die digital über bestimmte Programme ablaufen, also müssen die deklarierten Datentypen innerhalb der Programme bei Datentransformationsverfahren geprüft werden. Dieses Problem wird nicht weiter im Rahmen dieser Masterarbeit vertieft, da diese Thematik eine spezifische Angelegenheit der Informatik ist, dennoch gilt dies als ein zu berücksichtigender Faktor in der Umsetzung der Entscheidungsunterstützung. Diese Problematik wird in zwei Anforderung definiert, und zwar als Wertebereichanpassung und Datentypanpassung. Die Datentypanpassung ist in diesem Fall das spezifische Problem der Informatik.

In Abschnitt 2.3 wurden das Clustering und die Regression als Data-Mining-Methode vorgestellt und im Unterabschnitt 2.4.1 auch als Datentransformationsverfahren eingeordnet. Durch die Quellenvielfalt bei der Bearbeitung dieser Masterarbeit erfolgt die endgültige Einordung der Methoden in das Data Mining und werden dadurch nicht mehr als Datentransformationsverfahren betrachtet. Dennoch zeigt sich bei der Gegenüberstellung der Verfahren in den verschiedenen Prozessen, welche Anforderungen ebenfalls vorhanden sind. In der Datentransformation galten diese Verfahren nicht als Analyse Methoden, sondern als Methoden zur Verringerung des Datenbestands. Dies gilt nicht nur für das Clustering und die Regression, sondern auch für andere Methoden. Die Datentransformationsverfahren sollen im Data Mining für eine Entlastung der Rechenkapazitäten sorgen. Die Anforderung ist also die Reduzierung der Datenmenge.

Unabhängig von der Datenmenge ist die Datenkomplexität auch ein Faktor. In Abschnitt 4.1 wurde das Kriterium Dimension eingeführt. Dieses Kriterium thematisiert im mathematischen und im nominalen Sinne den Aufwand bzw. die Komplexität der vorhandenen Daten und der

Datentransformationsverfahren. Eine Anforderung des Data Mining an die Datentransformationsverfahren kann die Reduzierung der Komplexität sein. Wie im Unterabschnitt 2.4.1 vorgestellt, existieren Datentransformationsverfahren wie die Data-Cube-Aggregation, die als dimensionsreduzierende Verfahren gelten. Mit der Reduzierung der Komplexität soll die Analyse im Data Mining vereinfachter sein.

Die vorhandene Datenmenge wird durch den Anwender in der Produktion verwaltet. Dies bedeutet, dass er auf die Daten zugreifen und ggf. auch Veränderungen durchführen kann. Im Kontext der Herausforderungen des Data Mining ist die Datensicherheit ein Faktor, der berücksichtigt werden muss, aber in diesem Fall keine Bedeutung hat, da die Ausgangslage für den Anwender einen vollständigen Zugriff auf die Daten voraussetzt. Aufgrund dieser Annahme besteht hier keine Anforderung.

Eine weitere Herausforderung aus dem Abschnitt 2.2 ist die Verwaltung von Informationen aus verschiedenen Quellen. Dieser Aspekt wird ebenfalls durch die Skalenniveauanpassung und die Datentypanpassung bearbeitet. Außerdem beinhaltet diese Herausforderung auch die Datenspeicherung, die in Abschnitt 2.1 behandelt wurde. Eine Anforderung des Data Mining an die Datentransformation ist die sachlogische Strukturierung der Daten. Vielmehr ist diese Anforderung auch an die Datentransformation gerichtet und bezieht sich auf die Quelle der Daten. Mit der Quelle der Daten werden entweder Stammdaten oder Prozessdaten definiert. Im nächsten Abschnitt werden diese genauer beschrieben, da die Daten im Kontext dieser Arbeit aus der Automobilproduktion sind.

4.3 Randbedingungen der Automobilproduktion für Datentransformationsverfahren

In diesem Abschnitt werden die Randbedingungen der Automobilproduktion für die Datentransformationsverfahren thematisiert. Im Abschnitt 4.2 wurden zuletzt die Quellen der Daten besprochen. Wie im Unterabschnitt 2.5.2 vorgestellt, können die Daten aus der Automobilproduktion aus verschiedenen Quellen resultieren. Zu beachten sind Stamm-, Sensor- und Prozessdaten. Diese Daten können aus verschiedenen Prozessen entstehen oder sind wie Stammdaten gegeben. In Abschnitt 4.2 wurde die sachlogische Strukturierung der Daten bzgl. ihrer Speicherung vorgestellt. Im Unterabschnitt 2.5.2 wurde auf diese sachlogische Strukturierung in der Automobilproduktion ebenfalls eingegangen. Für die Daten aus der Automobilproduktion werden Datenbanken als geeignete Ablage bezeichnet. In diesen Datenbanken müssen dann logisch zusammengehörige Daten als konsistente Gruppe betrachtet werden. In der Automobilproduktion existieren verschiedene Fertigungstechnologien, die im Unterabschnitt 2.5.1 vorgestellt wurden. Aus diesen Fertigungstechnologien resultieren dann Daten, die aufgrund ihrer Quelle zusammenhängen sollten. Das bedeutet, dass Sensordaten, die aus dem Presswerk stammen miteinander gespeichert werden sollten und nicht z. B. mit den Sensordaten aus dem Karosserierohbau. Eine Ergänzung dazu sind die Stammdaten aus den Fertigungstechnologien. Diese sollten auch konsistent sein und für den Vergleich der aktuellen Prozessdaten bereitstehen. Als Randbedingung gilt somit die Betrachtung von Daten, die logisch zusammengehören und in diesem Fall aus relationalen Datenbanken sind.

Die Anforderung im Data Mining die Datenmenge zu reduzieren, muss auch im Kontext der Automobilproduktion betrachtet werden. Wie im Unterabschnitt 2.5.1 erläutert, hat die enorme Datenmenge auch einen Einfluss auf die Umweltbelastung. Die Datentransformation müssen kompatibel mit der Datenmenge sein und berücksichtigen, dass während der Transformation auch Rechenressourcen verbraucht werden. Mit der Datentransformation soll aber nicht nur die Datenmenge, sondern auch die Komplexität bearbeitet werden.

Ein weiterer Aspekt bei der Automobilproduktion ist, dass die Prozesse schnell ablaufen müssen. Je höher die Datenmenge und Komplexität, desto länger sind die notwendigen Verarbeitungszeiten. Wie in Abschnitt 2.5 thematisiert, sind Sensoren Bestandteile der Produktion und zuständig für die Umwandlung der Daten. Die Umwandlung der Signale erfolgt in Echtzeit, somit wird eine schnelle Datentransformation vorausgesetzt. In diesem Fall bedeutet das, dass höhere Datenmengen, die von den Sensoren resultieren länger verarbeitet werden. Wie in Abschnitt 4.2 thematisiert, ist der Aufwand ein Kriterium zur Anwendung von Datentransformationsverfahren. Einzelne Datentransformationsverfahren haben aufgrund ihrer eigenen Methodik innerhalb des Verfahrens mehr Rechenschritte. Eine Randbedingung aus der Automobilproduktion ist das Berücksichtigen des Aufwands bzgl. der Arbeit innerhalb der Sensoren. Aufgrund der resultierenden Datenmenge aus den Sensoren muss das gewählte Datentransformationsverfahren das vordefinierte Ziel erreichen und anhand der Methodik ausgewählt werden, damit geringe Verarbeitungszeiten vorhanden sind.

Des Weiteren ist die Datenqualität ein zu beachtender Aspekt. Während der Vorstellung der Kriterien in Abschnitt 4.1 fiel auf, dass die Datenqualitätsdimensionen aus dem Unterabschnitt 2.1.2 nicht allein auf den Prozess der Datentransformation übertragbar sind. Im Kontext der Automobilproduktion muss dennoch der Aspekt der Datenqualität berücksichtigt werden. Die vorgestellte Kategorie Kontextual ist mit dem Domänenbezug wieder von Bedeutung. Die Datenqualitätsdimension Datenmenge wurde, wie oben schon erwähnt als Randbedingung definiert. Die Datenqualitätsdimensionen Relevanz, Wert steigernd und Aktualität aus der Kategorie Kontextual sind ebenfalls in der Automobilproduktion zu beachten. Alle drei Datenqualitätsdimensionen sind für die Daten mit den erfassten Sensoren vorhanden. Neu erfasste Daten sind aktuell, für den jeweiligen Prozess relevant und haben das Ziel einer Wert Steigerung. Neben den vorhandenen Sensordaten müssen die anderen Daten auf diese Dimensionen geprüft werden. Am Anfang des Abschnitts 4.3 wurden Stammdaten genannt und ihr Auftreten in der Automobilproduktion thematisiert. Im Fall der Datenqualität stehen diese Stammdaten im Widerspruch zu der Datenqualitätsdimension, da sie nicht aktuell sein müssen. Die Automobilproduktion wird von dynamischen Prozessen beeinflusst und muss berücksichtigten, dass Stammdaten auf ältere Prozesse basieren. Als Randbedingung der Automobilproduktion für die Datentransformation resultiert, dass nur aktuelle bzw. neu erfasste Daten bearbeitet werden. Stammdaten können nur zum Vergleich benutzt werden, wenn der Prozess der Fertigung keine großen Veränderungen hatte. Die anderen Datenqualitätsdimensionen Relevanz und Wert steigernd werden mit der Aktualität ebenfalls berücksichtigt.

Eine weitere Randbedingung in der Automobilproduktion ist, dass bestimmte Prozesse ihre eigenen Daten definieren. Anwender können Vorgaben von Auftragsgebern erhalten, die den

Prozess beeinflussen und somit auch die Daten. Diese Vorgaben sind z. B. Parametereinstellungen. Je nach Einstellung resultieren unterschiedliche Daten, die dann verarbeitet werden müssen.

In den Kriterien zur Anwendung der Datentransformation und in den Anforderungen des Data Mining war die Veränderung des Skalenniveaus bzw. auch des Wertebereichs ein wesentlicher Faktor. In Abschnitt 3.4 wurde die Kategorisierung der Datentransformationsverfahren auch in diesem Kontext durchgeführt. Die verschiedenen Skalenniveaus geben an wie die Daten vorhanden sind. In Unterabschnitt 2.5.3 wurden verschiedene Anwendungsbeispiele vorgestellt, in denen Daten mit verschiedenen Skalenniveaus auftauchen. Die Skalenniveaus der Daten in der Automobilproduktion können als Randbedingung der Automobilproduktion definiert werden. Es muss beachtet werden, in welcher Skala die Daten entstehen und in welche Skala sie transformiert werden sollen. In einem der Anwendungsfälle wurden die Sensoren genauer ausgeführt und mögliche Ausgaben genannt. Wenn physikalische Werte gemessen werden, dann gibt der Sensor metrische Werte an, die kontinuierlich sind. Ergänzend dazu können die Sensoren auch nominale Daten ausgeben. Im Anwendungsfall werden Sensoren auch für Boolesche Operationen verwendet. Die Prüfung, ob Objekte vorhanden sind, wird durch eine nominale Ausgabe bestätigt. Verschiedene Skalenniveaus und Datentypen können in der Automobilindustrie auftauchen und müssen ebenfalls als Randbedingung der Automobilproduktion betrachtet werden.

Die Randbedingungen der Automobilproduktion wurden in diesem Abschnitt vorgestellt. Inhaltlich gibt es viele Überschneidungen zu den Kriterien aus Abschnitt 4.1 und zu den Anforderungen aus Abschnitt 4.2. Diese Überschneidungen bzw. Faktoren legen den Schwerpunkt in der zu konzipierenden Entscheidungsunterstützung fest.

4.4 Konzipierung der Entscheidungsunterstützung

In diesem Abschnitt wird das Hauptziel der Entwicklung einer Entscheidungsunterstützung zur Auswahl von Datentransformationsverfahren als Vorverarbeitung für das Data Mining in der Automobilproduktion angegangen. In Abschnitt 4.1 wurden Kriterien zur Anwendung der Datentransformationsverfahren erstellt, die nur auf die Methoden spezialisiert sind und weitere Einwirkungen vernachlässigen. In Abschnitt 4.2 und 4.3 wurden Anforderungen des Data Mining und die Randbedingungen der Automobilproduktion genannt. Die Anforderungen an das Data Mining wurden gestellt, da die Datentransformation als Vorverarbeitungsschritt für das Data Mining dient. In diesem Kontext wurden die Kriterien bzw. die Datentransformationsverfahren von ihrer alleinstehenden Betrachtung wieder in den Prozess des Data Mining eingeordnet. Ergänzend zu den Anforderungen an das Data Mining werden die Randbedingungen der Automobilproduktion beachtet, da diese die Domäne bzw. den Anwendungsbereich des Data Mining und der Datentransformation definiert. Zusätzlich müssen zur Konzipierung generelle Informationen bzgl. Entscheidungsunterstützungen berücksichtigt werden.

Dafür werden die Aspekte der Entscheidungsunterstützung aus dem Unterabschnitt 2.4.2 einbezogen. In dem Anwendungsrahmen dieser Arbeit wird eine Entscheidungsunterstützung konzipiert, die methodenorientiert ablaufen soll. Die Entscheidungsunterstützung soll nicht

überwachend dienen, sondern aktiv bei der Bearbeitung von komplexen Problemen zur Verfügung stehen. Ein weiterer Aspekt sind die Entscheidungen, die getroffen werden müssen. In der Entscheidungsunterstützung werden z. T. Entscheidungen auf objektive Fakten basieren und weitere auf subjektive Erfahrungen. Ein Beispiel für objektive Fakten ist, dass das vorhandene und das angestrebte Skalenniveau eindeutig zugeordnet werden kann. Als eine subjektive Erfahrung könnte die Einschätzung des Aufwands von ähnlichen Verfahren sein. Je nach Anwender muss die Einschätzung nicht nur vom Aufwand abhängen, sondern auch vom Wissensstand.

Die kognitiven Fähigkeiten, also auch die technische Ausbildung des Anwenders der Datentransformationsverfahren, muss ein bestimmtes Niveau entsprechen, um überhaupt Entscheidungen nachvollziehbar zu beschließen. Ergänzend zum Wissensstand bzw. zu den Erfahrungen des Anwenders hat die Risikobereitschaft auch einen Einfluss auf die Entscheidungen. Anwender, die eher dazu neigen Risiken einzugehen, streben schnellere Ergebnisse an, aber akzeptieren auch dafür Verluste. Ein wesentlicher Punkt sind auch Ziele in der Entscheidungsunterstützung. Im Unterabschnitt 2.4.2 wurde erwähnt, dass die Zieldefinition die Entscheidungsmöglichkeiten eingrenzt und diese nachvollziehbar gestaltet.

Diese Aspekte sind in der Konzipierung der Entscheidungsunterstützung für die Auswahl von Datentransformationsverfahren als Vorverarbeitung für Data Mining in der Automobilproduktion ebenfalls von Bedeutung. Um die Entscheidungsunterstützung zu konzipieren, werden in der folgenden Tabelle 7, die Kriterien aus Abschnitt 4.1, die Anforderungen aus Abschnitt 4.2 und die Randbedingungen der Automobilproduktion aus Abschnitt 4.3 gegenüber gestellt.

Tabelle 7: Gegenüberstellung der Kriterien zur Anwendung, Anforderungen des Data Mining und der Randbedingungen der Automobilproduktion

Kriterien zur Anwendung	**Anforderung des Data Mining**	**Randbedingungen der Automobilproduktion**
Dimension	Skalenniveauanpassung	Betrachtung sachlogisch strukturierten Daten
Ziel	Wertebereichanpassung	Reduzierung der Datenmenge & Komplexität
Aufwand	Datentypanpassung	Verschnellern der Prozesse (Betrachten des Aufwands)
Informationsgehalt	Reduzierung der Datenmenge	Aktualität der Daten (Veränderung der Prozesse)
Objektivität	Reduzierung der Komplexität	Prozessparameter und Einstellungen
Kombination	Sachlogische Speicherung der Daten	Skalenniveau von Daten

Kriterien zur Anwendung	Anforderung des Data Mining	Randbedingungen der Automobilproduktion
Vorverarbeitung		Datentypen

Für die Entscheidungsunterstützung wird ein Flussdiagramm erstellt, an dem sich der Anwender orientieren kann. Wie in Tabelle 7 zu erkennen ist, existieren Überschneidungen in den einzelnen Gruppen. Der Aspekt des Skalenniveaus ist in allen Gruppen vorhanden. Des Weiteren wird nicht nur das Skalenniveau in allen Gruppen thematisiert, sondern auch die Wertebereichanpassung. Beide Aspekte wurde schon in Abschnitt 3.4 als Hauptkategorien der Datentransformation vorgestellt. Der Punkt wurde mit dem Kriterium Ziel in Abschnitt 4.1 ausgearbeitet. Dieser Aspekt ist der Startpunkt in der konzipierten Entscheidungsunterstützung. Das Kriterium Aufwand hat Überschneidungen im Kontext der Reduzierung von Datenmenge und Komplexität. Dies soll in der Entscheidungsunterstützung durch die Optionen von verschiedenen Verfahren gegeben sein. Des Weiteren wird das Kriterium Informationsgehalt durch eine Bestätigung nach den Transformationen gegeben. Das Kriterium der Kombination soll durch eine Schleife eingebaut werden und somit die mehrfache Nutzung eines Datentransformationsverfahrens ermöglichen. Um die Entscheidungsunterstützung genauer zu untersuchen, wird in der folgenden Abbildung das fertige Konzept dargestellt:

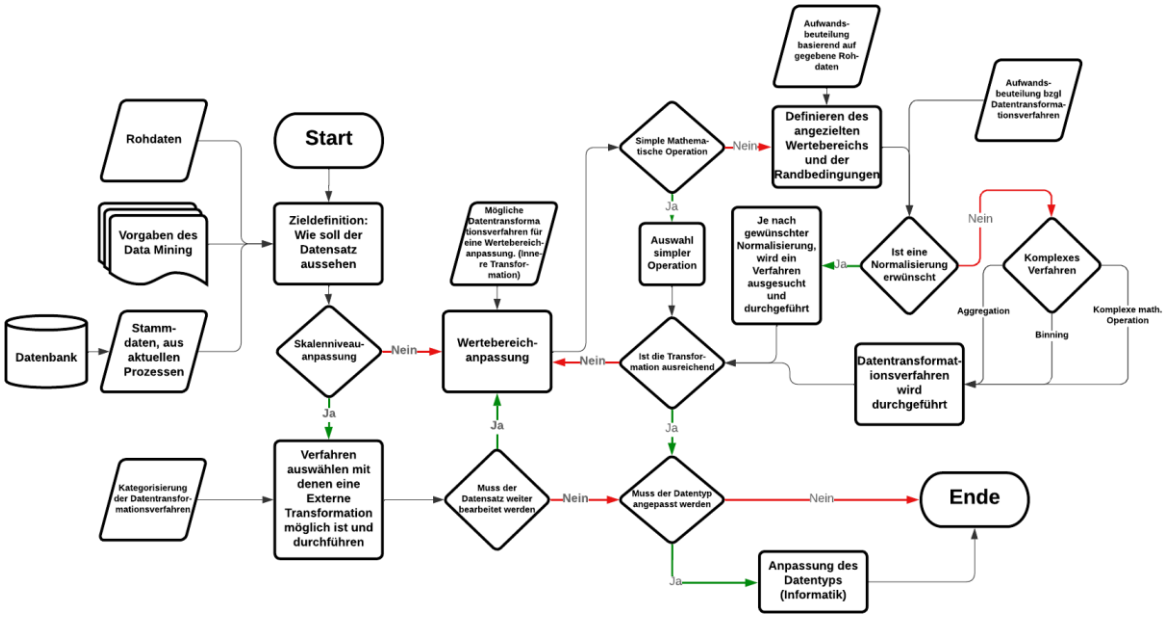


Abbildung 5: Entscheidungsunterstützung zur Auswahl von Datentransformationsverfahren als Vorverarbeitung für Data Mining in der Automobilproduktion

Wie in der Abbildung 5 zu erkennen ist, startet die Entscheidungsunterstützung mit der Zieldefinition. Die Zieldefinition hängt von mehreren Faktoren ab, wobei der Einfluss nicht gleichmäßig verteilt ist. Hauptsächlich bestimmt der Rohdatensatz und die folgende Analyse des Data Mining die Zieldefinition. Ergänzend werden Stammdaten aus den Datenbanken ausgegeben, um eine Referenz für den Anwender darzustellen.

Basierend darauf folgt in der Entscheidungsunterstützung die Entscheidung der Skalenniveauanpassung. Falls ja, dann wird mithilfe der Tabelle 5, in der die Kategorisierung der Datentransformationsverfahren vorhanden ist je nach vorliegendem Datensatz ein Verfahren für die Externe Transformation ausgewählt. Falls keine Skalenniveauanpassung erfolgt, wird an dem gegebenen Datensatz eine Wertebereichanpassung durchgeführt, bei der Verfahren aus der Kategorie Interne Transformation verwendet werden. An diesem Punkt folgen die ersten Entscheidungen des Anwenders, die hier mit der Zieldefinition begründet werden. Das bedeutet, dass in der Zieldefinition Vorgaben des Data Mining und Vergleichsdaten vorliegen müssen. Diese sollen den Anwender bei der Zieldefinition unterstützen. Der Anwender muss hier keine subjektive Entscheidung treffen, die auf seinen eigenen Erfahrungen basieren. Durch die Vorgaben des Data Mining, wie z. B. der Nennung der gewünschten Analysemethode, wird klar kommuniziert, wie der transformierte Datensatz auszusehen hat. Vergleichsdaten sollten auch nur betrachtet werden, wenn diese aus dem gleichen Prozess stammen.

Des Weiteren wird nach der Skalenniveauanpassung eine Wertebereichanpassung nicht ausgeschlossen. Dies bedeutet, dass nach der Externen auch eine Interne Transformation erfolgen kann. Falls dies nicht notwendig ist, muss nur noch im informatischen Sinne geprüft werden, ob die deklarierten Datentypen richtig sind. In diesem Fall wäre der erste Pfad in der Entscheidungsunterstützung beendet.

Der Prozess der Entscheidungsunterstützung läuft weiter, wenn eine Wertebereichanpassung benötigt wird. Als nächstes wird dann geprüft, ob eine simple mathematische Operation ausreicht. Diese simplen mathematischen Operationen wurden im Unterabschnitt 2.4.1 nicht detailliert ausgeführt, aber resultierten dennoch aus der systematischen Literaturrecherche. In Abschnitt 3.4 wurden die mathematischen Operationen erstmal als komplette Gruppe betrachtet und nicht nach Komplexität kategorisiert. In der Tabelle 5 der Kategorisierung der Datentransformationsverfahren wurden mathematischen Operationen anhand ihrer Dimension gruppiert. Mehrdimensionale Verfahren werden in diesem Kontext als komplexer betrachtet als die Gruppe Limitierte Dimension. Verfahren aus der Gruppe limitierte Dimension können im mathematischen Sinne eindimensional sein und dadurch einen geringen Aufwand besitzen.

Falls eine simple mathematische Operation durchgeführt wird, erfolgt als nächste Entscheidung, ob die Transformation ausreichend ist. Durch diese Entscheidung werden die Kriterien Informationsgehalt und Kombination geprüft. Der Anwender entscheidet hier, ob der Datensatz eine weitere Wertebereichanpassung benötigt. Falls dies nicht notwendig ist, wird wieder der Datentyp im Kontext der Informatik überprüft und der zweite Pfad der Entscheidungsunterstützung beendet.

Die Entscheidungsunterstützung wird fortgeführt, wenn eine simple mathematische Operation nicht ausreichend ist. In diesem Fall findet ein Prozess statt, bei dem Anwender klar definieren muss, wie der Wertebereich auszusehen hat. Des Weiteren können Randbedingungen definiert werden, die den Wertebereich oder generell den Datensatz bestimmen sollen. In diesem Prozess wird der gegebene Rohdatensatz genauer untersucht und aufgrund ihres Aufwands

beurteilt. Nachdem die Definitionen feststehen und der Anwender genau weiß, wie der Datensatz nach der Transformation auszusehen hat, muss er die verschiedenen Datentransformationsverfahren, die verwendet werden können, auch anhand ihres Aufwands beurteilen.

Der Anwender wird mithilfe der Entscheidungsunterstützung befragt, ob eine Normalisierung erwünscht ist. An dieser Stelle werden die verschiedenen Normalisierungsverfahren, die zur Wertebereichanpassung dienen, mit den restlichen komplexeren Verfahren getrennt. Falls eine Normalisierung stattfinden soll, wird der Prozess durchgeführt und der Anwender muss im Anschluss bewerten, ob die Transformation ausreichend ist. Wie oben erwähnt kann dann bei unzufriedenem Ergebnis eine weitere Datentransformation erfolgen. Falls der Anwender das Ergebnis akzeptiert, wird er nach der Datentypanpassung befragt und die Entscheidungsunterstützung endet.

Ein alternativer Weg entsteht, falls keine Normalisierung erwünscht ist. In diesem Fall muss der Anwender zwischen verschiedenen komplexen Verfahren entscheiden. Zur Auswahl stehen ihm komplexe mathematische Operationen, Aggregation- und Binningmethoden. An diesem Punkt muss der Anwender eine Gruppe an Verfahren wählen und dann innerhalb sich für eine Methode entscheiden. Nach der Entscheidung wird die Datentransformation durchgeführt und im Anschluss bewertet. Falls die Transformation ausreichend ist, wird mit der Datentypanpassung wieder die Entscheidungsunterstützung beendet.

Mit dem Schema aus Abbildung 5 kann der Anwender in der Automobilproduktion vor dem Data Mining eine Datentransformation durchführen. Das Schema dient als Anleitung und soll den Anwender während der Entscheidungsfindung zwischen den verschiedenen Kategorien der Datentransformationsverfahren unterstützen. Ergänzend zur Abbildung müssen noch bestimmte Aspekte berücksichtigt werden. Während der Nutzung dieser Anleitung ist ein bestimmtes Vorwissen des Anwenders vorausgesetzt. Des Weiteren sind die Tabellen 5 und 6 essenziell bei der Verwendung, da sie an bestimmten Stellen der Entscheidungsunterstützung benutzt werden. In der Wertebereichanpassung kommt hinzu, dass der Anwender nach der Definition seiner Vorstellung des transformierten Datensatzes innerhalb der verschiedenen Verfahren auch Auswahlmöglichkeiten hat, die speziell nach Anwendungsfall variieren. Außerdem sind im Schema weitere Datenvorverarbeitungsschritte nicht berücksichtigt worden, aber müssen in der Realität vor der Datentransformation geprüft werden.

Dies war die theoretische Vorstellung der konzipierten Entscheidungsunterstützung. Um diese zu prüfen, wird im nächsten Kapitel ein Fallbeispiel eingeführt und an diesem die Entscheidungsunterstützung angewendet.

5 Exemplarische Anwendung der Entscheidungsunterstützung

In diesem Kapitel wird die konzipierte Entscheidungsunterstützung aus Abschnitt 4.4 mithilfe eines Fallbeispiels validiert. Dafür wird im ersten Abschnitt ein Fallbeispiel aus der Automobilproduktion eingeführt. Bei der Einführung werden Randbedingungen und ggf. Vorgaben des Data Mining behandelt. In Abschnitt 5.2 wird an dem Datensatz des Fallbeispiels die Entscheidungsunterstützung verwendet, um ein mögliches Datentransformationsverfahren auszuwählen. Im letzten Abschnitt 5.3 werden mithilfe der Data-Mining-Analyse die Ergebnisse der Anwendung der Entscheidungsunterstützung vorgestellt und bestimmte Entscheidungen diskutiert.

5.1 Einführung eines Fallbeispiels

Für das Fallbeispiel wurde ein Datensatz von Kaggle (Missonnier 2024) ausgewählt, der Daten aus einem echten Herstellungsprozess simuliert. Dieser Datensatz wurde für die Beziehungen zwischen den verschiedenen Prozessparametern und der Produktqualität entwickelt. Der Datensatz hat zwei Prozessparameter, die Temperatur und der Druck, von denen die restlichen Parameter abhängen. Um den gemeinsamen Einfluss der beiden Prozessparameter darzustellen, sind im Datensatz die Spalten folgende Parameter ebenfalls vorhanden.

- Temperatur * Druck
- Material Schmelzmetrik
- Material Transformationsmetrik

Des Weiteren wird mit der Qualitätsbewertung die Gesamtqualität der Produkte wiedergegeben. Der Datensatz hat also sechs Spalten mit 3958 Zeilen. Auch wenn dieser Datensatz nur aus einer Simulation stammt, kann dieser für eine Auswertung der Produktion benutzt werden. Mit den gegebenen Parametern kann der Datensatz mit den Datensätzen aus der Automobilproduktion verglichen werden. Auch wenn alle Daten in dem gegebenen Datensatz zusammen gespeichert wurden, stammen die einzelnen Spalten nicht aus einem Prozess. Das bedeutet, dass die Qualitätsbewertung eigenständig betrachtet werden muss, im Gegensatz zu den anderen Parametern, die direkt aus den Prozessparametern Temperatur und Druck entstanden sind. Außerdem muss ergänzt werden, dass der Datensatz keine fehlenden Werte besitzt und somit in der ersten Betrachtung keine anderen Vorverarbeitungsschritte benötigt.

Um die Anwendung der Entscheidungsunterstützung anhand dieses Datensatzes vorzustellen, werden Szenarien erzeugt. Im ersten Szenario soll die Produktqualitätsbewertung aus dem Datensatz nicht nur durch einen Wert wiedergegeben werden, sondern durch Labels. Anwender und weitere Mitarbeiter sollen dadurch eine einfachere Zuordnung der Produktqualität erhalten und Ergebnisse besser interpretieren können.

Das zweite Szenario soll die Prozessparameter Temperatur und Druck untersuchen. In dem Kontext des Datensatzes basieren die weiteren Spalten auch auf den beiden Prozessparametern und diese sollen ebenfalls im Data Mining analysiert werden. Wie in Abschnitt 2.2 erwähnt

wurden, können die Methoden des Data Mining in beschreibende und vorhersagende Verfahren eingeteilt werden. In diesem Fall liegt ein Datensatz vor, der aus einer Simulation entstanden ist, bei dem das Beschreiben der resultierten Daten eher einen Mehrwert bringt. Als beschreibendes Verfahren wird das Clustering angewendet, um Gruppen zwischen den verschiedenen Datenpunkten zu erstellen. Im nächsten Abschnitt wird die konzipierte Entscheidungsunterstützung aus Abschnitt 4.4 benutzt, um die beiden Szenarien zu bearbeiten.

5.2 Anwendung der Entscheidungsunterstützung

Für das erste Szenario müssen zuerst Anforderungen des Data Mining definiert werden. Wie in Abschnitt 2.2 erwähnt, sollen im Data Mining die Daten analysiert und im Nachhinein interpretiert werden. Das Ziel ist in diesem Fall die Datentransformation der Spalte Qualitätsbewertung. Im Datensatz ist zu erkennen, dass die Werte der Qualitätsbewertung in einem Intervall zwischen 1 und 100 liegen. In diesem Fall steht die 1 für die schlechteste und die 100 für die beste Produktqualität. In dem Datensatz sind dazwischen ebenfalls Werte vorhanden und müssen vom Anwender in der Produktion eingeschätzt werden. In solchen Skalen ist die Bewertung der Qualität immer mit Problemen verbunden, da klare Grenzen der Produktqualität nicht definiert wurden. Um diese Problematik zu lösen, wird die numerische Darstellung der Produktqualität thematisiert. In der folgenden Abbildung wird der Pfad der Entscheidungsunterstützung durchlaufen:

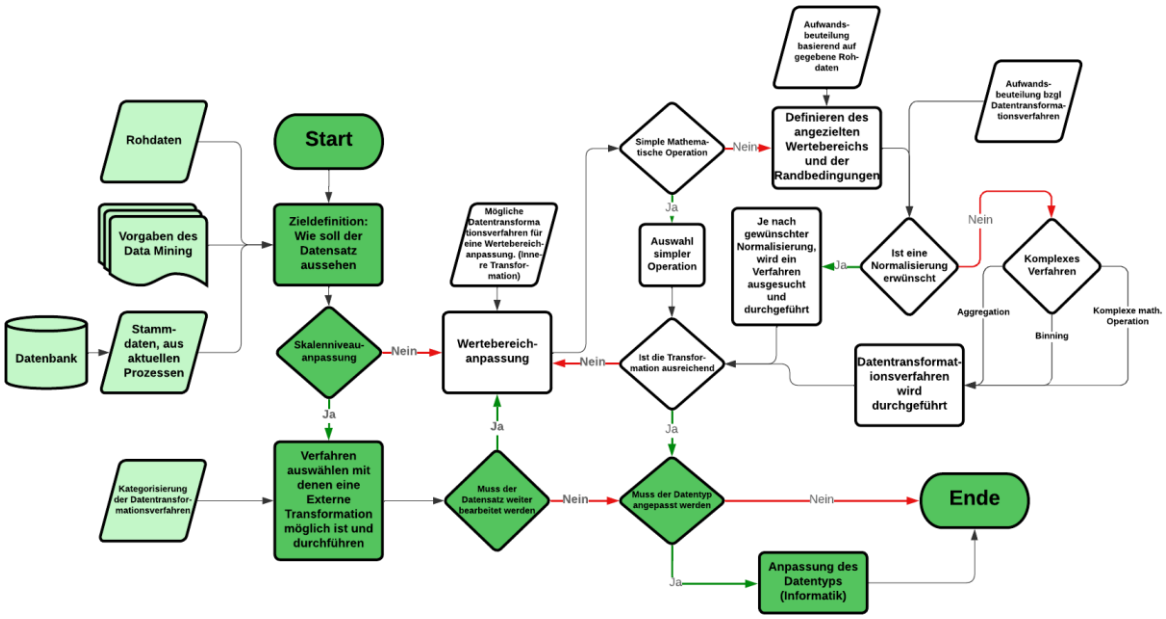


Abbildung 6: Pfad der Entscheidungsunterstützung für das erste Szenario

Wie in der Abbildung 6 zu erkennen ist, wurde der Pfad der Entscheidungsunterstützung in grün und Dateninputs in hellgrün markiert. Es ist zusehen, dass in dem ersten Szenario eine Skalenniveauanpassung vorgenommen wird, da die numerischen Werte transformiert werden sollen. Wie oben erwähnt ist die Zieldefinition, dass kontinuierliche Werte in ein Intervall trans-

formiert werden sollen. Das bedeutet, dass nach der Entscheidung der Skalenniveauanpassung ein Verfahren ausgewählt werden soll, mit denen eine externe Datentransformation möglich ist. Hierzu wird ein Ausschnitt aus der Kategorisierung der Datentransformationsverfahren aus Abschnitt 3.4 betrachtet:

Tabelle 8: Externe Transformationen aus der Kategorisierung der Datentransformationsverfahren

Externe Transformation
Gramian-Angular-Field-Encoding
One-Hot-Encoding
Discretization
Discretization-Entropy-Based
Discretization-By-Clustering
Discretization-By-Binning
Discretization-By-Histogram
Generalization

Die Tabelle 8 ist ein Ausschnitt aus der Kategorisierung der Datentransformationsverfahren und beinhaltet die externen Datentransformationsverfahren. Der Anwender muss nach der Entscheidung, dass eine Skalenniveauanpassung erfolgen soll, diesen Tabellenausschnitt genauer untersuchen. Zuerst muss der Anwender wissen, welches Skalenniveau vorhanden ist und welches als Ziel definiert wurde. Darauf basierend werden die Verfahren nach ihrer Methodik ausgewählt. In diesem Szenario sollen numerische Werte durch Labels ersetzt werden. Die ersten beiden Verfahren behandeln andere Anwendungsfälle und sind nicht für den gegebenen Datensatz passend.

Des Weiteren muss der Anwender die verschiedenen Verfahren der Discretization und die Generalization untersuchen. Wie im Unterabschnitt 2.4.1 erwähnt, bezieht sich die Generalization auf die Darstellung der Attribute durch höhere Ebenen. Die Discretization thematisiert ebenfalls die Darstellung der Attribute mit höheren Hierarchien, aber bezieht sich direkt auf das Skalenniveau und verdeutlicht die Transformation von numerischen Werten in das nominale Skalenniveau. In diesem Anwendungskontext ist die Discretization passender, da in der Methode genau die notwendige Skalenniveauanpassung definiert wurde. Um die Discretization durchzuführen, muss der Anwender Vorgaben des Data Mining und Randbedingungen der Automobilproduktion beachten.

Es können zwei Unterscheidungen getroffen werden. Entweder werden die numerischen Werte in eine boolesche Ausgabe transformiert und geben an, ob die Produktqualität ausreicht oder ungenügend ist. Die zweite Option ist, dass die Produktqualität mit der Skala Niedrig, Mittel und Hoch kategorisiert wird. Dabei muss der Anwender hier eine Entscheidung treffen,

die abhängig von den Produktionsvorgaben ist. In diesem Fall wird die Discretization durchgeführt und die Skala mit Niedrig, Mittel und Hoch übernommen.

Für die Skala der Discretization müssen Schwellenwerte definiert werden. Wie erwähnt wurden in den Daten die Qualitätsbewertung zwischen 1 und 100 angegeben. Die mathematisch gleich großen Intervalle würden in diesem Fall nicht passend sein, da die Bewertung Hoch der Skala mit dem Wert 67 zu niedrig ist. Vielmehr muss die Skala so angepasst werden, dass der Schwellenwert für die Bewertung Niedrig den größten Bereich bzw. das größte Intervall abdeckt. In der Produktion müssen Vorgaben klar für die Anwender definiert werden, wenn Produkte aufgrund von Mängeln nicht ausreichend sind. Der Sinn der Skala ist die Steigerung des Verständnisses und somit auch das Erkennen von mangelnder Qualität. Mit der Bewertung Niedrig werden somit Produkte als nicht ausreichend gekennzeichnet. Das zweite Intervall mit der Bewertung Mittel wird für Produkte verwendet, die Mängel besitzen, aber dennoch nach einer Nachbearbeitung verwendet werden können. Dies bedeutet, dass die Produkte als ausreichend gelten, aber aufgrund von Defekten trotzdem nochmal genauer untersucht werden sollten. Die letzte Kategorie mit der Bewertung Hoch sind dann Produkte, die eine ausgezeichnete Qualität besitzen und das kleinste Intervall definieren. Folgende Discretization wird ausgeführt:

- Von 1 bis 75,99 = Niedrig.
- Von 76 bis 90,99 = Mittel.
- Von 91 bis 100 = Hoch.

Nachdem die Skalenniveauanpassung durchgeführt wurde, wird im Schema der Entscheidungsunterstützung untersucht, ob für den Datensatz noch eine weitere Bearbeitung notwendig ist. Im Fall des Szenarios ist es nicht notwendig und dadurch führt der Pfad zur Entscheidung, ob der Datentyp angepasst werden muss. Aufgrund der Transformation des Skalenniveaus von numerischen in nominalen Daten, ändert sich der Standarddatentyp von Integer zu Char. Somit endet der Pfad für das erste Szenario mit einer Skalenniveauanpassung.

Für das zweite Szenario wird wieder ein Pfad in der Entscheidungsunterstützung durchlaufen und dafür in der folgenden Abbildung dargestellt:

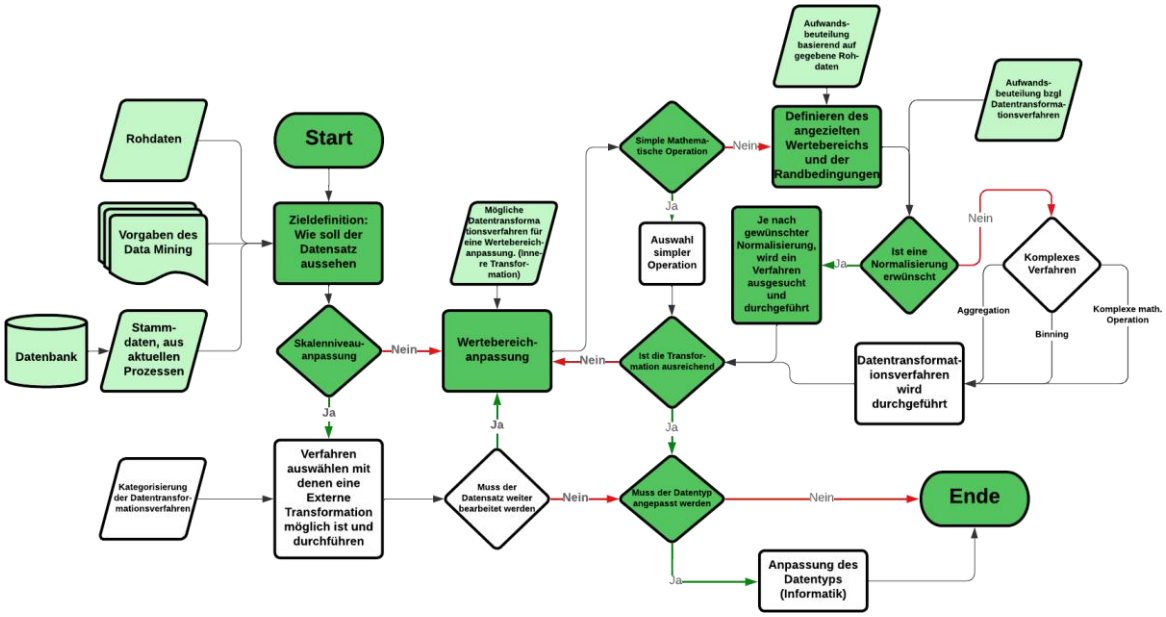


Abbildung 7: Pfad der Entscheidungsunterstützung für das zweite Szenario

Wie im ersten Szenario wirken in der Zieldefinition verschiedene Aspekte des Data Mining. In diesem Fall ist im Data Mining ein Clustering geplant und dadurch wird die Zieldefinition beeinflusst. Außerdem werden die Sensordaten bestehend aus Temperatur und Druckwerten dann mit den Stammdaten aus den gleichen Prozessen verglichen. In der Zieldefinition wurde festgelegt, dass für die Werte der Temperatur und des Drucks eine Transformation benötigt wird, da die Wertebereiche der Parameter unterschiedlich sind.

Nachdem definiert wurde, wie der Datensatz sein soll, wird als erste Entscheidung geprüft, ob eine Skalenniveauanpassung erforderlich ist. Aufgrund der Zieldefinition weiß der Anwender, dass nicht das Skalenniveau, sondern der Wertebereich angepasst werden muss. Im Prozessschritt der Wertebereichanpassung stehen dem Anwender die internen Transformationsverfahren aus der Tabelle 5 zur Verfügung. Der Anwender muss in diesem Schritt entscheiden, ob eine simple mathematische Operation durchgeführt werden soll oder eher ein komplexeres Verfahren benötigt wird. In diesem Fall wird aufgrund des benötigten Wertebereichs ein komplexeres Verfahren verlangt. Der Anwender muss an diesem Punkt seinen Wertebereich genauer definieren und mögliche Randbedingungen beachten. Dazu betrachtet er wieder den gegebenen Datensatz und muss dabei vorhandene Komplikationen erkennen. In diesem Szenario muss der Anwender die Werte der Temperatur und des Drucks vergleichen und ein Verfahren unter Berücksichtigung des möglichen Aufwands auswählen.

Als nächstes wird vom Anwender entschieden, ob der gewünschte Wertebereich durch eine Normalization erreicht werden kann. Für das Clustering der Temperatur- und Druckwerte ist eine Normalization geeignet, da die verschiedenen Wertebereiche angepasst werden. Somit können die verschiedenen Werte der Parameter vereinfacht in einem Bereich dargestellt und in schnellerer Zeit verarbeitet werden. Nach der Entscheidung muss sich der Anwender je nach Datensatz und bekannten Normalisierungsmethoden für ein Fall entscheiden. Wie im

Unterabschnitt 2.4.1 erklärt, ist die Datenverteilung ebenfalls entscheidend für die Verfahrensauswahl. Dies führt dazu, dass die Werte der Temperatur und des Drucks ebenfalls auf Ausreißer untersucht werden müssten.

In diesem Fall sind keine Ausreißer vorhanden und dadurch auch die Min-Max-Normalization und die Z-Scale-Normalization möglich. Des Weiteren werden alle Werte der Temperatur und des Drucks untersucht. Nach der Entscheidung des Verfahrens und der Durchführung muss der Anwender bestimmen, ob die Datentransformation ausreichend ist. Entweder akzeptiert der Anwender die Ergebnisse der Normalization oder führt eine weitere Wertebereichanpassung durch. In diesem Fall ist die Normalization ausreichend und der Datentyp wird überprüft. Aufgrund der Werte ist keine Datentypanpassung erforderlich, da die Veränderung der Wertebereiche nicht sehr groß ist.

Somit endet der Pfad der Entscheidungsunterstützung für das zweite Szenario. Um die Entscheidungsunterstützung zu bewerten, wird im nächsten Abschnitt die Auswirkung der getroffenen Entscheidungen betrachtet und mithilfe der Analyse des Data Mining untersucht.

5.3 Ergebnisse

In diesem Abschnitt werden die Entscheidungen aus den beiden Szenarien, die mithilfe der entwickelten Entscheidungsunterstützung getroffen wurden, untersucht. Im ersten Szenario fand eine Skalenniveauanpassung statt. Dabei wurden numerische Werte der Qualitätsbewertung in drei Kategorien diskretisiert. Aufgrund der vorhandenen Qualitätsstandards der Automobilproduktion war die Einteilung der Kategorien nicht gleichmäßig. Um die hohen Erwartungen in der Produktion bzw. auch im Wettbewerb einzuhalten, wurde ein Schwellenwert angegeben, der nicht ausreichende Produkte differenziert. Damit die Datentransformation aus dem ersten Szenario bzw. die Auswahl des Verfahrens bewertet werden kann, wird eine Data-Mining-Methode durchgeführt.

Um das erste Szenario zu bearbeiten, werden die Werte der Qualitätsbewertung anhand der Intervalle aus dem letzten Abschnitt diskretisiert. Für die Umsetzung wurde Python 3.13.2 mit der Bibliothek Pandas 2.2.3 verwendet, die für Datenanalyse bzw. auch Datenvorverarbeitung benutzt wird. Die Vorteile an der genutzten Bibliothek sind die Funktionen zur Datenlesung und Bearbeitung von Tabellen. Für die Vereinfachung wurde die Spalte der Qualitätsbewertung separiert und als einzelner Datensatz gespeichert. Des Weiteren wurde die Datei in das CSV-Format exportiert. Der Algorithmus ist im Anhang unter B.1 wiederzufinden.

Durch die Discretization entstehen zwei wichtige Aspekte für die Anwender in der Produktion. Wie in Abschnitt 2.2 erwähnt, werden im letzten Schritt des Data-Mining-Prozesses Ergebnisse ausgewertet und interpretiert. Kognitive Fähigkeiten sind nicht nur während der Entscheidungsfindung von Bedeutung, sondern auch in der Nachbereitung. Durch das Ersetzen der numerischen Werte, wird das Verständnis der Bewertung für die Anwender verbessert. Außerdem kann die Discretization auch die Verwendung einer Data-Mining-Methode nützlicher ge-

stalten. In diesem Fall soll mithilfe einer Assoziationsanalyse im folgenden Abschnitt der Zusammenhang zwischen Temperatur, Druck und Qualitätsbewertung festgestellt werden. Damit die Assoziationsanalyse durchgeführt werden kann, werden die numerischen Werte der Temperatur und des Drucks ebenfalls diskretisiert. Der Algorithmus aus dem Anhang B.1 wird mit bestimmten Modifikationen auch für die Parameter Temperatur und Druck verwendet. Im Gegensatz zur Qualitätsbewertung werden gleichgroße Intervalle des gesamten Prozesses erstellt. Für die Temperatur folgen diese Intervalle:

- Von 100,01 bis 166,66°C = Niedrig.
- Von 166.67 bis 233,33°C = Mittel.
- Von 233,34 bis 300°C = Hoch.

Der gleiche Prozess wird auch für den Parameter Druck durchgeführt und dadurch entstehen folgende Intervalle:

- Von 5 bis 11,66 kPa = Niedrig.
- Von 11,67 bis 18,33 = Mittel.
- Von 18,34 bis 25 = Hoch.

Diese Intervalle wurden in Python umgesetzt, damit die numerischen Werte der Parameter auch diskretisiert werden. Zur Vereinfachung der Leistung wurden die Spalten, die nicht notwendig sind für beide Szenarien entfernt. In dem erweiterten Algorithmus, der im Anhang unter B.2 zu sehen ist, wurden die beiden Prozessparameter und die Qualitätsbewertung mit den angegebenen Intervallen diskretisiert. In den Datensätzen ist zu erkennen, dass die vortransformierten Daten ebenfalls als Spalten vorhanden sind und direkt miteinander verglichen werden können. An diesem Punkt kann der Anwender erkennen, ob die Datentransformation erfolgreich abgelaufen ist. Da der Algorithmus eine CSV-Datei erzeugt, wurde dann überprüft, ob die Datentypen auch stimmen. Die automatische Datentyperkennung von Excel wurde in diesem Schritt ignoriert, da bestimmte Werte falsch erkannt wurden.

Bei der Analyse des resultierenden Datensatzes fiel ein inhaltlicher Fehler auf, der durch die Formatierung im Algorithmus entstand. Inhaltlich zusammenhängende Spalten wurden nicht zeilenweise betrachtet. Das bedeutet, dass die Spalten ihre Abhängigkeit untereinander verloren haben. Während der Fehlerüberprüfung wurde entschieden, dass die Discretization innerhalb der Excel-Datei erfolgen soll. Mithilfe der Wenn-Funktion in Excel konnten die vorbestimmten Intervalle eingesetzt und durch die Kategorien ersetzt werden. Einerseits ist diese Methode einfacher in der Umsetzung als die in Python und andererseits konnten dabei die Spalten kontrolliert werden. Die Wenn-Funktion für die Qualitätsbewertung lautet:

=WENN(C1<76;"Niedrig“; WENN(C1<91;"Mittel";"Hoch"))

Wie zu erkennen ist, wurden die Intervalle durch die Schwellenwerte 76 und 91 erzeugt. Analog zur Qualitätsbewertung wurde die Wenn-Funktion auch für die Temperatur und den Druck umgesetzt. Außerdem werden die Begriffe Niedrig, Mittel und Hoch, je nach Spalte mit dem Anfangsbuchstaben des Parameters gekennzeichnet. Diese Vorgehensweise soll dabei helfen, die diskreten Begriffe ohne Spalten zuzuordnen. Im Fall der Temperatur wird der Begriff

Niedrig zu $T_{Niedrig}$. Folgende Tabelle ist ein Ausschnitt aus den Ergebnissen der Discretization in Excel:

Tabelle 9: Ausschnitt aus den Ergebnissen der Discretization

Temperature (°C)	Pressure (kPa)	Quality Rating	Temperature diskretisiert	Pressure diskretisiert	Quality Rating diskretisiert
209,7627008	8,050855372	99,99997052	T_{Mittel}	$P_{Niedrig}$	Q_{Hoch}
243,0378733	15,81206837	99,98570341	T_{Hoch}	P_{Mittel}	Q_{Hoch}
220,5526752	7,843130045	99,99975761	T_{Mittel}	$P_{Niedrig}$	Q_{Hoch}

In der Tabelle wurden die ersten drei Ergebniszeilen dargestellt. Wie zu erkennen ist, sind die drei linken Spalten die gegebenen Parameter des Datensatzes. Mit diesen Spalten wurden die drei rechten Spalten durch die Anwendung der Wenn-Funktion mithilfe der vorgestellten Schwellenwerte erzeugt. Das Resultat ist eine Tabelle, die alle Einträge der Temperatur, des Drucks und der Qualitätsbewertung in numerischer und diskretisierter Form beinhaltet.

Nachdem die Discretization durchgeführt wurde, konnte die Assoziationsanalyse starten. Dafür wurden in Python neben der Pandas Bibliothek auch noch die Mlxtend Bibliothek mit der Version 0.23.4 eingeführt, die als Machine-Learning-Extension definiert wird. Die Mlxtend Bibliothek bietet die Funktionen der Assoziationsanalyse und somit auch den Apriori Algorithmus an. Im Anhang unter B.3 ist der Code für die Assoziationsanalyse vermerkt. Wie zu erkennen ist, wurde der Apriori Algorithmus aus Abschnitt 2.3 verwendet, um die Assoziationsanalyse durchzuführen. Das erste Ziel war die Ermittlung der Frequent Itemsets, die im Code mit einem Supportwert ermittelt werden. Der Supportwert wurde in Abschnitt 2.3 als Wahrscheinlichkeit des zusammen auftreten der Objekte bezeichnet. Als Supportwert wurde zuerst 0,3 gewählt, damit viele Muster bzw. Frequent Itemsets gefunden werden. Das bedeutet, die Objekte treten in mindestens 30% der Zeilen zusammen auf. Ergänzt zum Supportwert wird die Konfidenz ermittelt. In Abschnitt 2.3 wurde erwähnt, dass die Konfidenz den Zusammenhang der Objekte an gibt. Um die Assoziationsregeln aufzustellen, wurde im Code ein Konfidenzwert von 70% ausgewählt. Folgende Ergebnisse resultieren für den Supportwert mit dem diskretisierten Datensatz:

Tabelle 10: Ergebnisse der Frequent Itemsets mit einem Support von 30%

Supportwert	Frequent Itemset
0,908011	Q_{Hoch}
0,316907	P_{Hoch}
0,341168	P_{Mittel}

Supportwert	Frequent Itemset
0,341926	$P_{Niedrig}$
0,336366	T_{Hoch}
0,324994	T_{Mittel}
0,338640	$T_{Niedrig}$
0,307809	Q_{Hoch}, P_{Mittel}
0,313369	Q_{Hoch}, $P_{Niedrig}$
0,324994	Q_{Hoch}, T_{Mittel}
0,338640	Q_{Hoch}, $T_{Niedrig}$

Wie in der Tabelle 10 zu erkennen ist, besteht der Datensatz zu ca. 90% aus Produkten, die in der Qualitätsbewertung Q_{Hoch} zugeordnet bekommen haben. Mithilfe der Tabelle und dem gewählten Supportwert von 0,3 sind vier Frequent Itemsets mit zwei Items zu erkennen. Im Vergleich der vier Frequent Itemsets mit jeweils zwei Items fiel auf, dass die Bewertung $T_{Niedrig}$ am häufigsten mit Q_{Hoch} auftritt. Dennoch ist der Unterschied des Supportwerts nicht signifikant. Außerdem wurde identifiziert, dass mit dem Supportwert von 0,3 T_{Hoch} und P_{Hoch} nicht ausgegeben wurden und eine geringere Assoziation zu einer hohen Qualitätsbewertung besitzen.

Im nächsten Schritt der Assoziationsanalyse wird die Konfidenz ermittelt, mit dem der Zusammenhang der Frequent Itemsets bestimmt wird. Folgende Tabelle resultiert für die relevanten Frequent Itemsets:

Tabelle 11: Assoziationsregeln mit Konfidenzwerten

Assoziationsregel	Konfidenzwert
P_{Mittel} -> Q_{Hoch}	0.9022222222222221
$P_{Niedrig}$ -> Q_{Hoch}	0.9164818920916482
T_{Mittel} -> Q_{Hoch}	1.0
$T_{Niedrig}$ -> Q_{Hoch}	1.0

In der Tabelle 11 ist zu erkennen, dass alle vier Beziehungen einen hohen Konfidenzwert besitzen und dadurch einen starken Zusammenhang haben. Insbesondere sind die Temperaturskalen T_{Mittel} und $T_{Niedrig}$ mit einem Konfidenzwert von 1 sehr ausschlaggebend. Dieser Wert bedeutet, dass jedes Mal, wenn $T_{Niedrig}$ oder T_{Mittel} erscheinen, Q_{Hoch} daraus resultiert. Die Assoziation von diesen beiden Frequent Itemsets sind somit auf der höchsten Stufe. Der Einfluss des Drucks ist ebenfalls vorhanden, kann aber nicht mit der Beziehung von der Temperatur

zur Qualität mithalten. In der Produktion bedeutet das, dass bei Qualitätsproblemen eine Kontrolle der Temperatur erfolgen muss. Dennoch steht das T_{Hoch} nicht automatisch für eine schlechte Qualität der Produkte. In diesem Szenario wurden die numerischen Werte der Temperatur und des Drucks gleichgroß aufgeteilt. In realen Fällen können Schwellenwerte für die Prozessparameter bekannt sein, bei denen Anwender wissen, dass die Produkte Qualität verlieren. Mit diesem Wissen können Prozesse der Produktion genauer untersucht und für die Discretization passendere Intervalle ausgewählt werden.

Das erste Szenario wurde mit der Assoziationsanalyse beendet und hat die Entscheidungsfindung in Abschnitt 5.2 validiert. Das Hauptziel in der Entscheidungsfindung war die Änderung des Skalenniveaus, um das Verständnis bzw. die Interpretierbarkeit der Parameter für den Anwender zu verbessern. Die Discretization wurde in diesem Abschnitt nicht nur für die Qualität, sondern auch für die Prozessparameter durchgeführt. Dis diskreten Begriffe waren für die Assoziationsanalyse von großer Bedeutung, um die Zusammenhänge der Parameter in vereinfachter Form zu sehen und das Verständnis zu erhöhen. Somit wurde die Entscheidungsfindung bzw. das Datentransformationsverfahren hauptsächlich durch die Data-Mining-Methode beeinflusst.

Als nächstes wird das zweite Szenario bearbeitet. Wie in Abschnitt 5.2 erwähnt, soll eine Normalization der Prozessparameter stattfinden, um danach ein Clustering durchzuführen. Damit die Vorteile der Normalization gesehen werden sollen, wird das Clustering zuerst ohne Normalization dargestellt. Um die Cluster zu erschaffen, wird wieder Python verwendet und mit den Bibliotheken Matplotlib 3.10.0 und Scikit Learn 1.6.1 erweitert. Ersteres ermöglicht die Visualisierung in Python. Mit Scikit Learn können verschiedene Machine-Learning-Algorithmen angewendet werden. In Python werden beide Bibliotheken in Kombination mit Pandas benutzt, um das zweite Szenario zu bearbeiten. Ergänzend zu den Prozessparametern wird für das Clustering auch die Qualitätsbewertung miteinbezogen. Dabei wird die Spalte der Qualitätsbewertung nicht für das Clustering selbst, sondern für die Analyse benutzt. Wie in Abschnitt 2.3 erklärt, wird der k-Means-Algorithmus verwendet, der durch die Bibliothek von Scikit Learn gegeben ist. In diesem Fall wurde die Anzahl der Cluster auf drei gesetzt und die Anzahl der maximalen Iterationen auf 50. Für die Visualisierung in Python wurden dann die Plot Funktionen von Matplotlib verwendet. Dies ermöglichte eine saubere Abgrenzung der Cluster. In der folgenden Abbildung werden zuerst die Punkte der Prozessparameter nach ihrer Qualität dargestellt ohne die Berücksichtigung der Cluster:

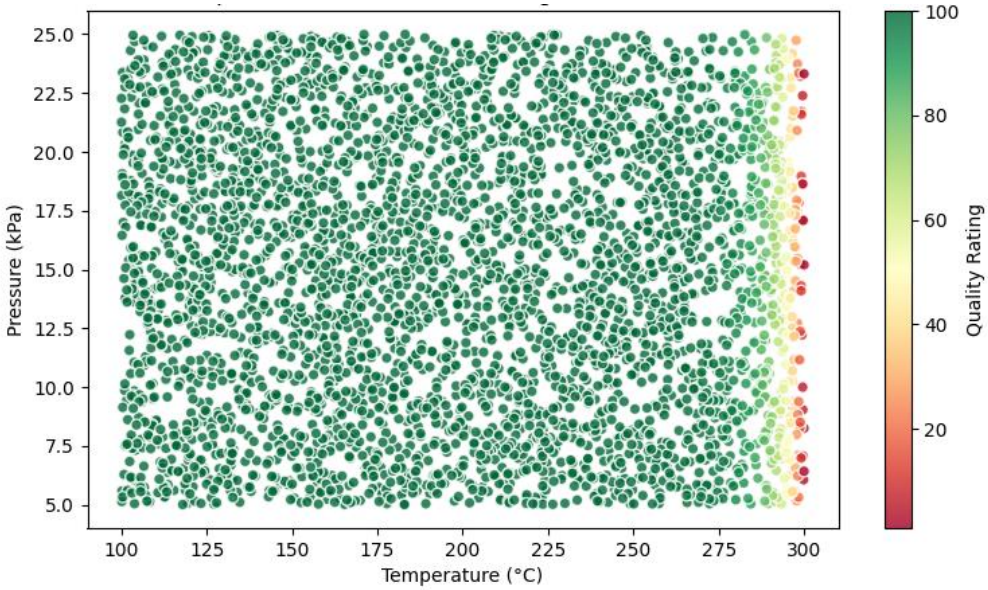


Abbildung 8: Datenpunkte der Prozessparameter mit farblicher Visualisierung der Qualität

In der Abbildung 8 ist der Plot zu erkennen, der aus dem Algorithmus aus dem Anhang B.4 resultiert ist. Die Datenpunkte wurden farblich nach ihrer Qualität visualisiert und zeigen, dass die Qualität stark von der Temperatur abhängt, so wie es die Assoziationsanalyse ebenfalls festgestellt hat. In diesem Fall sieht es so aus, als ob der Druck keinen Einfluss auf die Qualität hat. Im nächsten Schritt sollen die Datenpunkte mit den Clustern dargestellt werden, ohne dass die Normalization durchgeführt wird. Dafür werden folgende Zeilen aus dem Algorithmus B.4 angepasst:

- scatter = matplt.scatter(Datensatz['Temperature (°C)'], Datensatz['Pressure (kPa)'], c=Datensatz['Quality Rating'], cmap='RdYlGn', edgecolors='white', alpha=0.8)
- matplt.title("Datenpunkte mit Qualitätsbwertung ohne Normalization")
- matplt.colorbar(label="Quality Rating")

Die letzte Zeile mit der Colorbar wird komplett entfernt, da in diesem Fall die Qualität keinen Einfluss auf das Clustering hat. Die beiden anderen Zeilen werden folgenderweise verändert:

- scatter = matplt.scatter(Datensatz['Temperature (°C)'], Datensatz['Pressure (kPa)'], c=Datensatz['Cluster'], cmap='Greys', edgecolors='black', alpha=0.8)
- matplt.title("Clustering ohne Normalization")

In der Scatter Funktion wurden die farblichen Aspekte angepasst. Die neue Abbildung wird nicht nach der Qualität farblich gekennzeichnet, sondern anhand ihrer Cluster Zugehörigkeit. Die untere Zeile veranschaulicht nur die Veränderung des Titels der Abbildung. Durch die Änderungen resultiert folgende Abbildung:

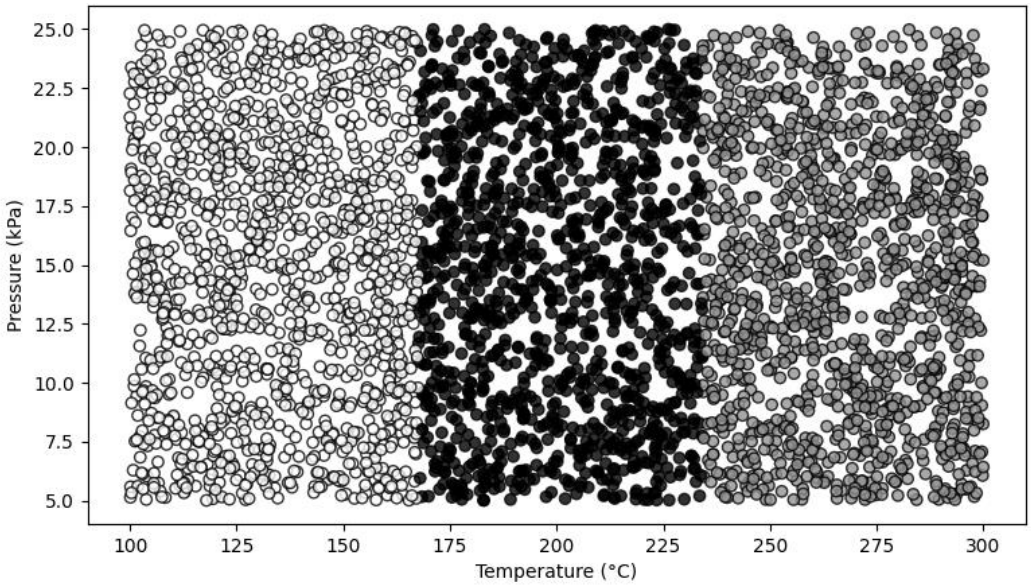


Abbildung 9: Clustering der Datenpunkte ohne Normalization

Wie in der Abbildung 9 zu erkennen ist, sind drei Cluster entstanden, die farblich unterschiedlich dargestellt werden. Es fällt dabei auf, dass die Cluster nur abhängig von der Temperatur visualisiert wurden und keinen Bezug zum Druck haben. Ein möglicher Grund ist die fehlende Normalization, da im k-Means-Algorithmus mit Distanzen Cluster berechnet werden. Die Unterschiede der Wertebereiche können die Clusterbildung beeinflussen und werden im nächsten Schritt durch die Normalization bearbeitet. Im Anhang unter B.5 ist der Algorithmus für das Clustering mit Normalization vermerkt und erzeugt folgende Abbildung:

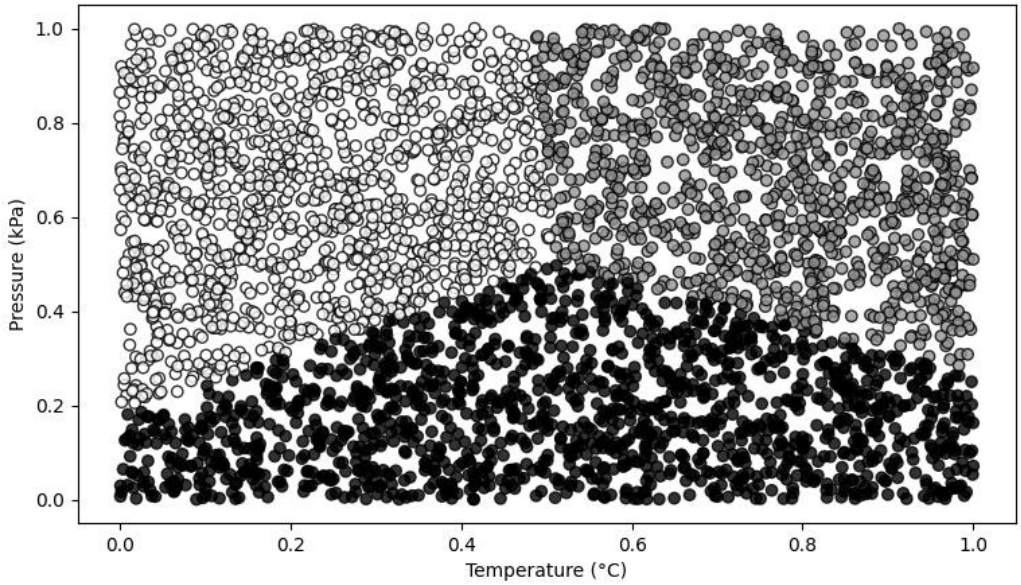


Abbildung 10: Clustering der Prozessparameter mit Normalization

In der Abbildung 10 ist zu erkennen, dass die Cluster nicht nur von der Temperatur abhängen, sondern nach der Normalization auch vom Druck. Im Algorithmus wurde die vordefinierte Funktion für die Min-Max-Normalization verwendet und deshalb sind die Wertebereiche für

beide Prozessparameter zwischen 0 und 1. In dieser Betrachtung können die drei Cluster anhand der Prozesszustände eingeordnet werden. Dabei muss beachtet werden, dass die Gruppe mit den weißen Datenpunkten keine Produkte mit schlechter Qualitätsbewertung enthält. Diese Gruppe entstand aus Prozessen, bei denen niedrige bis mittlere Temperatur und mittlerer bis hoher Druck vorhanden waren. Im Gegensatz dazu sind in den anderen beiden Clustern Datenpunkte vorhanden, die eine schlechte Produktqualität vorweisen. Um diese Datenpunkte wieder zu prüfen, wurde die Darstellung aus der Abbildung 8 mit den normalisierten Wertebereichen der Prozessparameter angepasst. Dafür wurde der Algorithmus B.4 mit den Normalization Funktionen erweitert. Es resultiert in der folgenden Abbildung die Darstellung der Qualität mit normalisierten Wertebereichen:

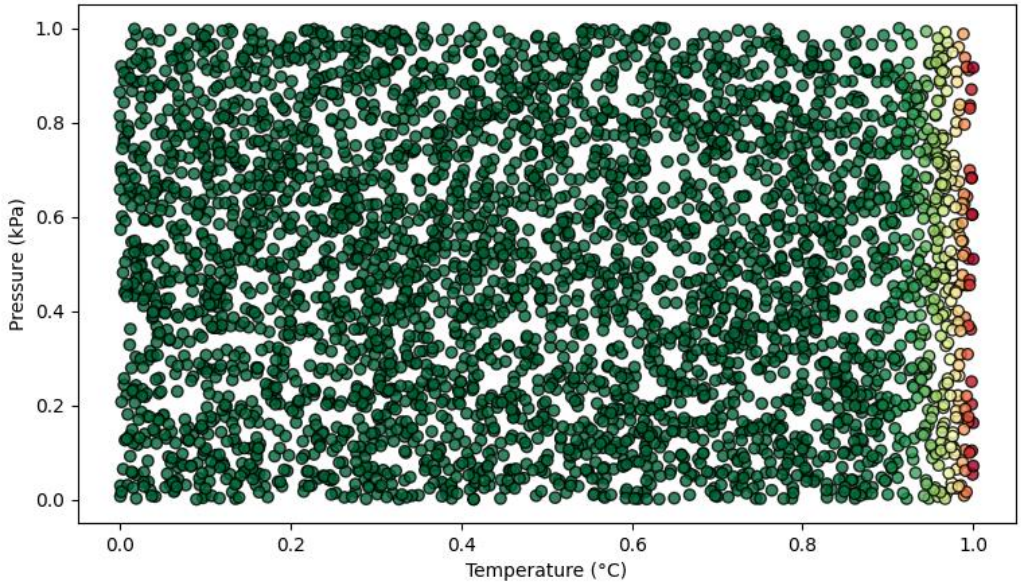


Abbildung 11: Datenpunkte in normalisierten Wertebereichen der Prozessparameter mit farblicher Visualisierung der Qualität

In der Abbildung 11 werden die Datenpunkte wieder anhand ihrer Qualität durch eine farbliche Darstellung gekennzeichnet. Wenn Abbildung 10 und 11 verglichen werden, fällt auf, dass die Datenpunkte immer noch in den beiden Clustern vorhanden sind. Vielmehr können diese Datenpunkte auch nach der Normalization hauptsächlich der Temperatur zugeordnet werden.

Das bedeutet, dass die erzeugten Cluster vor und nach der Normalization reininhaltlich nicht vom Druck abhängig sind. Die neuen Distanzen zwischen den Datenpunkten führen zu anderen Clustern, aber haben keine Prozessmuster hervorgehoben, die eine Druckabhängigkeit nachweisen. Mit den Erkenntnissen aus den beiden Data-Mining-Methoden erfolgt, dass die Temperatur der wichtigste Faktor in der Qualität der Produkte ist und der Druck somit in der Prozessoptimierung vernachlässigt werden kann. In der Assoziationsanalyse wurden niedriger und mittlerer Druck mit hoher Qualität als Frequent Itemset definiert. In dem ersten Szenario wurden Frequent Itemsets unter einem Support von 0,3 nicht weiter untersucht. Möglicherweise wurde deshalb das Frequent Itemset hoher Druck mit hoher Qualität vernachlässigt. Durch die erneute Durchführung der Assoziationsanalyse mit einem neuen Supportwert von 0,25 resultiert, dass P_{Hoch} mit Q_{Hoch} zu 0,28 korreliert. In der Analyse wurde durch die Wahl des

Supportwerts diese Erkenntnis nicht miteinbezogen und durch das Clustering bzw. durch die Darstellung der Datenpunkte entdeckt.

Somit wurde die Analyse für beide Szenarien beendet und führt zu den folgenden Informationen für die Entscheidungsunterstützung. Beide Szenarien wurden für einen Datensatz innerhalb dieser Arbeit entwickelt. Im Vorfeld stand nicht fest, welche Data-Mining-Methode angewendet werden soll und wurde deshalb ohne Vorgaben für die Szenarien geschaffen. Die Data-Mining-Methode mit den Parametern, die untersucht werden sollen, bestimmen größtenteils die Entscheidung zwischen den Datentransformationsverfahren. Für den Anwender ist es von großer Bedeutung genau zu wissen, welche Methode des Data-Mining benutzt wird. Ergänzend dazu ist es sinnvoll die Kategorisierung der Datentransformationsverfahren, die in der Entscheidungsunterstützung angewendet wurden, mit Informationen bzgl. der Verfahren zu erweitern. Im Fall des ersten Szenarios wäre es für die Anwender hilfreich, wenn neben den Verfahren vermerkt wird, welche Skalenniveauanpassung vorgenommen wird. Dies ermöglicht eine schnellere Entscheidungsfindung und senkt den Anspruch an Vorwissen.

Des Weiteren ist die Information des vorausgesetzten Skalenniveau in den verschiedenen Data-Mining-Methoden ebenfalls eine Vorgabe, die den Anwender bei der Entscheidungsfindung unterstützt. Das bedeutet, dass sobald der Anwender eine klare Zielvorstellung hat, die auch die spezifische Nutzung einer Methode des Data Mining beinhaltet, sollte er das Skalenniveau der vorhandenen Daten und die angestrebte Transformation verstehen. Die Skalenniveauanpassung ist dann erfolgreich, wenn die Rohdaten mit dem neuen Skalenniveau im Data Mining funktionieren.

Während der Bearbeitung des zweiten Szenarios wurden die inneren Transformationen aus der Kategorisierung der Datentransformationsverfahren verwendet. In diesem Fall wurde das Clustering als Ziel angegeben und darauf basierend Entscheidungen getroffen. Die Entscheidung, dass eine Normalization erfolgen soll, wurde schon bei der ersten Betrachtung der vorhandenen Daten festgestellt. In dem Szenario wurde der Pfad der Entscheidungsunterstützung mit der Normalization beendet. Während der Durchführung des Clustering wurde dann bemerkt, dass die Daten noch eine weitere Datentransformation benötigt hätten. Der Datensatz hat viele Datenpunkte, wodurch die Bearbeitung in Python länger dauert. In dem Fall hätte eine Aggregation von einer bestimmten Anzahl an Werten, die Datenmenge verringert. Vor allem fiel dabei das Verhältnis von Datenpunkten unterschiedlicher Qualität auf, da der Datensatz hauptsächlich Produkte wiedergibt, die eine hohe Qualität haben.

Dennoch wurden beide Szenarien mithilfe der Entscheidungsunterstützung bearbeitet und entsprechend für das Data Mining vorbereitet. Eine generelle Bestätigung der Entscheidungsunterstützung kann an dieser Stelle noch nicht erfolgen, da diese nur anhand eines Beispieldatensatzes mit erschaffenen Szenarien validiert wurde. Es existieren verschiedene Aspekte, die noch diskutiert werden müssen, um ein Fazit zur Entscheidungsunterstützung zu ziehen.

6 Diskussion und Fazit

Die Entwicklung einer Entscheidungsunterstützung zur Auswahl von Datentransformationsverfahren als Vorverarbeitung für das Data Mining in der Automobilproduktion war das Ziel dieser Arbeit. In der Literatur sind verschiedene Datentransformationsverfahren vorhanden, die in unterschiedlichen Domänen angewendet werden. Durch die verschiedenen Anwendungsbereiche wurden Datentransformationsverfahren weiter entwickelt. Für einen Anwender innerhalb einer Domäne, wie in diesem Fall in der Automobilproduktion, ist die Auswahl des passenden Datentransformationsverfahren mit Herausforderungen verbunden. Dabei ist die Schwierigkeit nicht nur von den Datentransformationsverfahren abhängig, sondern auch von den Vorgaben des Data Mining und der Domäne. Auch wenn in der Literatur die Anwendung der Datentransformation in der Automobilproduktion thematisiert wird, existiert für die Auswahl zwischen den verschiedenen Verfahren kein Leitfaden. Um das Fazit dieser Arbeit zu ziehen, werden die einzelnen Schritte zur Zielerreichung diskutiert.

Innerhalb dieser Arbeit wurden Datentransformationsverfahren mit einer systematischen Literaturrecherche ermittelt, die auf Begrifflichkeiten aus dem Stand der Technik basieren. Dafür wurde im Stand der Technik eine generelle Übersicht von den bewährten Datentransformationsverfahren vorgestellt und in den Kontext des Data-Mining-Prozesses gesetzt. Im Prozess des Data Mining sind die Datenvorverarbeitung und somit auch die Datentransformationsverfahren vor der Analyse zu berücksichtigende Faktoren. Aufgrund des Fokus wurde in dieser Arbeit das Data Mining mit bestimmten Analysemethoden untersucht, wodurch schon im Stand der Technik Komplikationen aufgrund der Verfahren Clustering und Regression entstanden. In verwendeten Quellen wurden die beiden Data-Mining-Methoden auch als Datentransformationsverfahren vorgestellt. Vor der finalen Durchführung der systematischen Literaturrecherche wurde diese Definition verworfen, da in den ersten Entwicklungen des Suchstrings auffiel, dass damit größtenteils die Analyse im Data Mining assoziiert wird. Die Definitionen der Datentransformationsbegriffe aus dem Stand der Technik wurden als erstes Teilziel umfasst und wurden durch die Ermöglichung der systematischen Literaturrecherche erreicht.

Aus dem finalen Suchstring der systematischen Literaturrecherche resultierten Ergebnisse, die aufgrund der Anzahl gering erschienen. Die geringe Anzahl an Ergebnissen ist aufgrund des restriktiven Suchstrings entstanden und wurde durch eine Rückwärtssuche ergänzt. Dafür wurden die Quellen der Publikationen aus der systematischen Literaturrecherche betrachtet und gefundene Datentransformationsverfahren hinzugefügt. In der generellen Durchführung der systematischen Literaturrecherche wurden nicht viele Publikationen ausgeschlossen, sodass die finale Anzahl an Quellen gepasst hat. Das zweite Teilziel thematisierte die Kategorisierung von Datentransformationsverfahren, die mit den Resultaten der systematischen Literaturrecherche erfüllt wurde. Die Datentransformationsverfahren wurden in der Kategorisierung in die Gruppen Externe oder Interne Transformationen unterteilt. Diese Unterteilung hat sich zu diesem Zeitpunkt nur auf die Anwendung der Datentransformationsverfahren bezogen und hat den ersten Schritt zur Entscheidungsunterstützung geschaffen.

Im nächsten Schritt wurden in der Arbeit Kriterien zur Anwendung der Datentransformationsverfahren bestimmt. Für die Erstellung dieser Kriterien wurden Aspekte aus dem Stand der Technik und aus der Kategorisierung der Datentransformationsverfahren entnommen. Ein umstrittener Punkt dabei war die Datenqualität bzw. die Datenqualitätsdimensionen, da das Nutzen dieser für eine Datentransformation komplex ist. Die eingeführten Datenqualitätsdimensionen aus 2.1.2 beziehen sich auf generelle Datenqualität und haben nicht den Kontext der Datentransformation als Vorverarbeitung des Data Mining. Eine vollständige Übernahme der Datenqualitätsdimensionen war somit nicht möglich und wurden deshalb modifiziert. Einige der Kriterien zu Anwendung der Datentransformationsverfahren wurden durch Anpassungen der Definition von Datenqualitätsdimensionen bestimmt. Das Kriterium Informationsgehalt z. B. resultiert aus der Datenqualitätsdimension Genauigkeit und hat einen Nutzen in der Entscheidungsunterstützung.

Nach den Kriterien wurden die Vorgaben des Data Mining thematisiert. Das Data Mining stellte eine ähnliche Betrachtung wie in den Kriterien dar. Das vorhandene Skalenniveau und der Wertebereich des Datensatzes muss für die folgende Analyse im Data Mining vorbereitet werden. Wieder einmal wurde auf die externe und interne Transformation eingegangen, die in diesem Fall unter Skalenniveauanpassung definiert wurden. Des Weiteren wurden allgemeine Verarbeitungen der Daten, wie die Reduzierung der Datenmenge und der Komplexität, als Anforderungen gestellt. Ein kritischer Punkt hierbei ist die Speicherung der Daten. Im Unterabschnitt 2.1.1 wurden verschiedene Datenebenen vorgestellt, die die Strukturierung der Daten untersuchen. Im Kontext dieser Arbeit wurden Datenbanken statt Big Data fokussiert, da Datenbanken eine sachlogische Strukturierung in überschaubarer Menge liefern. Big Data als Anwendungsbereich wäre problematischer bzgl. Rechenleistung und auch aufgrund der Datenvielfalt. Außerdem werden Datenbanken als klassische Datenebene in der Automobilproduktion betrachtet.

Somit war die Datenebene auch ein Faktor der Automobilproduktion, die in der Entwicklung der Entscheidungsunterstützung betrachtet werden muss. Nach den Vorgaben des Data Mining wurden die Randbedingungen der Automobilproduktion thematisiert. Diese Randbedingungen der Automobilproduktion wurden im Rahmen der Datenwissenschaft betrachtet und waren ähnlich zu den Vorgaben des Data Mining. Spezifische Prozesse der Automobilproduktion erzeugen Daten und müssen danach transformiert werden. Ergänzend zu den Vorgaben des Data Mining werden hierbei die Entstehungsprozesse bzw. der Ursprung der Daten zum Thema gebracht. Wenn aus dem gleichen Prozess Stammdaten existieren, ist es für den Anwender simpler die Datentransformation durchzuführen, da ein Beispiel vorhanden ist. Deshalb werden die Vorgänge durch die Betrachtung der Beispielprozesse verschnellert. Wie in den Kriterien und in den Vorgaben des Data Mining ist auch hier das vorhandene Skalenniveau ein wichtiger Aspekt. In diesem Fall kann der Anwender durch die Sensorik und der zugehörigen Codierung das Skalenniveau beeinflussen und muss dann im Anschluss je nach Data-Mining-Methode eine Anpassung durchführen.

Mit den Kriterien zur Anwendung der Datentransformationsverfahren, den Vorgaben des Data Mining und den Randbedingungen der Automobilproduktion wurde die Konzipierung der Entscheidungsunterstützung angezielt. Diese drei Faktoren zu verbinden, umfasste das dritte Teilziel.

Mit dem Fallbeispiel in Kapitel 5 erfolgte die exemplarische Validierung der Entscheidungsunterstützung und somit das letzte Teilziel. Hierbei gab es einige Punkte die genauer betrachtet werden müssen, bevor das Fazit gezogen werden kann. Durch die Szenarien fiel auf, dass die Data-Mining-Methode als hauptsächlicher Faktor die Entscheidungsfindung beeinflusst. Die Zieldefinition in der Entscheidungsunterstützung muss die Analyse des Data Mining beinhalten bzw. die zu nutzende Methode. Mit diesem Wissen können vor allem Skalenniveauanpassungen objektiver durchgeführt werden. Dabei ist es von Bedeutung, dass der Anwender wissen muss, welches Skalenniveau der vorhandene Datensatz hat und welches Skalenniveau in der Data-Mining-Methode vorausgesetzt wird. Durch diesen Schritt kann der Anwender viele der Datentransformationsverfahren ausschließen und ein Verfahren auswählen, mit dem er dieses Ziel erreichen kann.

Die Subjektivität des Anwenders wird viel mehr in den Entscheidungen der Wertebereichanpassungen erforderlich. Zwischen den komplexen Verfahren, den simplen mathematischen Operationen und den Methoden der Normalization muss der Anwender basierend auf den gegebenen Datensatz nochmal seine Zielvorstellung definieren, die in den meisten Fällen von der Erfahrung abhängt. Die simplen mathematischen Operationen wurden in der systematischen Literaturrecherche auch mit den Begrifflichkeiten der Datentransformation entdeckt und wurden im Konzept der Entscheidungsunterstützung berücksichtigt. Dennoch werden diese differenziert betrachtet und werden in den meisten Fällen nicht mit einer Datentransformation assoziiert. Andererseits sind die komplexen Verfahren und die Methoden der Normalization im Stand der Technik unter Datentransformationsverfahren vorgestellt worden und daher eher im Fokus in der Wertebereichanpassung. Innerhalb dieser Entscheidungen muss der Anwender, wenn er keine Vorgaben hat, frei zwischen mehreren Verfahren auswählen. In der exemplarischen Validierung gab es für das zweite Szenario deshalb nicht nur eine Lösung bzw. richtige Entscheidung, sondern mehrere. In dem Abschnitt 5.2 wurde dieser Aspekt beachtet und mehrere Verfahren der Normalization als verwendbar definiert.

Außerdem fiel nach der Durchführung der Discretization aus dem ersten Szenario auf, dass Datentransformationen auch innerhalb der Umsetzung der Algorithmen wieder auftauchen können. Der Pfad der Entscheidungsunterstützung hat nur den inhaltlichen Aspekt des Data Mining berücksichtigt und nicht die Umsetzung in der Informatik. Mit der letzten Entscheidung, ob der Datentyp angepasst werden muss, wurde nur das Ergebnis der Daten berücksichtigt. In der Durchführung der Assoziationsanalyse mussten die diskreten Werte im Algorithmus wieder transformiert werden. Für die Bearbeitung war eine binäre Darstellung der diskreten Werte essenziell. Diese Aspekte werden in der Entscheidungsunterstützung nicht betrachtet, aber können in der Aufwandsbeurteilung von Informatikern berücksichtigt werden, wenn das Data Mining erfolgt.

Des Weiteren wurde in der Entscheidungsunterstützung nur die Datentransformation behandelt und keine anderen Vorverarbeitungsschritte. In den Kriterien wurde dargelegt, dass andere Vorverarbeitungsschritte ebenfalls relevant für die Datentransformation sind und deshalb beachtet werden sollten. Dennoch gehört die Durchführung eines anderen Datenvorverarbeitungsschritts nicht in die Entscheidungsunterstützung und muss entweder davor oder danach erfolgen. In den Szenarien wurde dies mithilfe der Python Algorithmen bearbeitet und der Datensatz z. B. auf fehlende Werte geprüft. Ein weiterer Aspekt ist, dass nach der Durchführung der Methode des Data Mining bemerkt werden kann, dass die Datentransformation nicht ausreichend war. In diesem Fall kann die Entscheidungsunterstützung wieder ab dem Punkt „Ist die Transformation ausreichend“ bearbeitet und die Informationen aus der durchgeführten Data-Mining-Methode benutzt werden.

Mit den Ergebnissen der exemplarischen Validierung lässt sich schließlich das Fazit ziehen. Das Hauptziel dieser Arbeit war die Entwicklung einer Entscheidungsunterstützung zur Auswahl der Datentransformationsverfahren als Vorverarbeitung des Data Mining in der Automobilproduktion, da dieser Bereich in der Literatur fehlt. Die entwickelte Entscheidungsunterstützung hat mit dem Fallbeispiel gezeigt, dass diese an Produktionsdaten angewendet werden kann. Dabei ist es von Bedeutung, dass der Anwender ein grundlegendes Wissen bzgl. Datentransformationsverfahren und darüber hinaus auch vom Data Mining hat. Diese beiden Faktoren beeinflussen die Zieldefinition, die als wichtigster Schritt in der Entscheidungsunterstützung bezeichnet werden kann. Die erstellten Kriterien zur Anwendung von Datentransformationsverfahren und die Kategorisierung der Datentransformationsverfahren haben für die Entwicklung der Entscheidungsunterstützung die Grundlage geschaffen und sind für das Verständnis des Anwenders zu berücksichtigen. Des Weiteren ist die Kategorisierung der Datentransformationsverfahren nicht nur auf den jetzigen Inhalt beschränkt und kann mit der Zeit erweitert werden. Zusammenfassend kann gesagt werden, dass die Entscheidungsunterstützung mit der Kategorisierung der Datentransformationsverfahren und dem Wissen der Kriterien zur Anwendung von Datentransformationsverfahren in der Automobilproduktion genutzt werden kann.

7 Zusammenfassung und Ausblick

Mit der vorliegenden Arbeit wurde das Ziel verfolgt, eine Entscheidungsunterstützung zur Auswahl von Datentransformationsverfahren als Vorverarbeitung für Data Mining in der Automobilproduktion zu entwickeln. Dafür wurden in Kapitel 2 Datentransformationsverfahren vorgestellt und in den Kontext des Data Mining gesetzt. Ergänzend dazu wurden die Entscheidungsunterstützung allgemein und die Automobilproduktion als Anwendungsdomäne eingeleitet.

Mit dem Wissen aus dem Stand der Technik erfolgte in Kapitel 3 die systematische Literaturrecherche zu den verschiedenen Datentransformationsverfahren. Die resultierenden Datentransformation wurden kategorisiert und haben somit den ersten Schritt zur Konzipierung der Entscheidungsunterstützung geschaffen. Mit der Kategorisierung der Datentransformationsverfahren wurden in Kapitel 4 die Kriterien zur Anwendung gestellt, die sich auf inhaltliche Aspekte fokussieren. Anschließend wurden die Vorgaben des Data Mining definiert, die sich auf die Analysemethoden und generell auf die Daten bezogen. Im nächsten Schritt wurden die Randbedingungen der Automobilproduktion thematisiert, die sich größtenteils ebenfalls auf die Daten stützten. An diesem Punkt wurden die drei relevanten Aspekte zur Entwicklung der Entscheidungsunterstützung fertig gestellt und in den Zusammenhang gebracht. Ergänzend dazu wurden die Informationen zur Entscheidungsunterstützung aus dem Stand der Technik übernommen, damit die Konzipierung erfolgen konnte. Die Konzipierung resultierte in einer grafischen Darstellung des Entscheidungsprozesses, wodurch ein Anwender den Ablauf der Entscheidungen nachverfolgen kann.

Mit dem Schema der Entscheidungsunterstützung zur Auswahl von Datentransformationsverfahren wurde ein Fallbeispiel in Kapitel 5 bearbeitet. Die Bearbeitung hatte das Ziel, die entwickelte Entscheidungsunterstützung exemplarisch zu validieren. Durch die Ergebnisse der exemplarischen Validierung wurden bestimmte Aspekte der Entscheidungsunterstützung hervorgehoben und im nächsten Kapitel diskutiert. Als Fazit folgte, dass die Entscheidungsunterstützung in ihrer jetzigen Form anwendbar ist und durch Optimierungen vereinfacht werden kann.

Ein Ausblick ist die Optimierung der Entscheidungsunterstützung durch das Ergänzen der Skalenniveaus an den Datentransformations- und Data-Mining-Verfahren. Außerdem kann die Kategorisierung durch neue Datentransformationsverfahren erweitert werden. In weiteren Forschungsarbeiten könnte die Entscheidungsunterstützung in den Wertebereichanpassungen vertieft werden und z. B. innerhalb der Verfahren der Normalization Kriterien bzw. Entscheidungsmöglichkeiten hervorheben. Des Weiteren könnte dieser Ansatz nicht nur für die Normalization, sondern auch für die Aggregation fokussiert werden, da beide Verfahren in der Literatur häufig vertreten sind.

8 Literaturverzeichnis

Abdel-Karim, Benjamin M. (2022): Data Science. Wiesbaden: Springer Fachmedien Wiesbaden.

Abou Eddahab, Fatima-Zahra; Horvath, Imre (2018): What Does Data Analytics Offer for Extracting Knowledge from Middle-of-Life Product Data? In: *25th ISPE International Conference on Transdisciplinary Engineering 2018*, S. 1102–1111.

Aggarwal, Charu C. (2017): Data Mining. Cham: Springer International Publishing.

Alasadi, Suad A.; Bhaya, Wesam S. (2017): Review of Data Preprocessing Techniques in Data Mining. In: *Journal of Engineering and Applied Sciences* (12), S. 4102–4107.

Alpar, Paul; Alt, Rainer; Bensberg, Frank; Weimann, Peter (2019): Anwendungsorientierte Wirtschaftsinformatik. Wiesbaden: Springer Fachmedien Wiesbaden.

Altuhaifa, Fatimah Abdulazim; Win, Khin Than; Su, Guoxin (2023): Predicting lung cancer survival based on clinical data using machine learning: A review. In: *Computers in biology and medicine* 165, S. 107338. DOI: 10.1016/j.compbiomed.2023.107338.

Alvarez-Coello, Daniel; Wilms, Daniel; Bekan, Adnan; Marx Gómez, Jorge (2021): Towards a Data-Centric Architecture in the Automotive Industry. In: *Procedia Computer Science* 181, S. 658–663. DOI: 10.1016/j.procs.2021.01.215.

Al-Wesabi, Fahd N.; Malibari, Areej A.; Mustafa Hilal, Anwer; NEMRI, Nadhem; Kumar, Anil; Gupta, Deepak (2022): Intelligent ensemble of voting based solid fuel classification model for energy harvesting from agricultural residues. In: *Sustainable Energy Technologies and Assessments* 52, S. 102040. DOI: 10.1016/j.seta.2022.102040.

Ammal, S. Meenakshi; Kathiresh, M.; Neelaveni, R. (2021): Artificial Intelligence and Sensor Technology in the Automotive Industry: An Overview. In: M. Kathiresh und R. Neelaveni (Hg.): Automotive Embedded Systems. Cham: Springer International Publishing (EAI/Springer Innovations in Communication and Computing), S. 145–164.

Anbarasi, J. Deepa; Radha, V. (2023): Ridge Regularized Imputed Scaled Clipping Normalization Based Pre-processing for Marine Weather Forecasting. In: *IJCTE* 15 (2), S. 68–75. DOI: 10.7763/IJCTE.2023.V15.1333.

Anusha, Yamijala; Visalakshi, R.; Srinivas, Konda (2023): Imbalanced data classification using improved synthetic minority over-sampling technique. In: *MGS* 19 (2), S. 117–131. DOI: 10.3233/MGS-230007.

Avila Cobos, Francisco; Alquicira-Hernandez, José; Powell, Joseph E.; Mestdagh, Pieter; Preter, Katleen de (2020): Benchmarking of cell type deconvolution pipelines for transcriptomics data. In: *Nature communications* 11 (1), S. 5650. DOI: 10.1038/s41467-020-19015-1.

Backhaus, Klaus; Erichson, Bernd; Gensler, Sonja; Weiber, Rolf; Weiber, Thomas (2023): Multivariate Analysemethoden. Wiesbaden: Springer Fachmedien Wiesbaden.

Badshah, Afzal; Daud, Ali; Alharbey, Riad; Banjar, Ameen; Bukhari, Amal; Alshemaimri, Bader (2024): Big data applications: overview, challenges and future. In: *Artif Intell Rev* 57 (11). DOI: 10.1007/s10462-024-10938-5.

Baumöl, U.; Österle, H.; Winter, R.; Jung, Reinhard; Winter, Robert (2000): Data Warehousing Strategie. Berlin, Heidelberg: Springer Berlin Heidelberg.

Beckschulte, Sebastian; Huebser, Louis; Günther, Robin; Kaden, Miriam; Schmitt, Robert H. (2023): Fehlerdatenaufnahme in der manuellen Montage: Informationsbedarfsanalyse für die Fehleranalyse und -abstellung im Fehlermanagement. In: Sophie Gröger (Hg.): Nachhaltiges Qualitätsdatenmanagement. Wiesbaden: Springer Fachmedien Wiesbaden, S. 1–20.

Bewong, Michael; Wondoh, John; Kwashie, Selasi; Liu, Jixue; Liu, Lin; Li, Jiuyong et al. (2023): DATM: A Novel Data Agnostic Topic Modeling Technique With Improved Effectiveness for Both Short and Long Text. In: *IEEE Access* 11, S. 32826–32841. DOI: 10.1109/ACCESS.2023.3262653.

Bissantz, Nicolas; Hagedorn, Jürgen (2009): Data Mining. In: *Bus. Inf. Syst. Eng.* 1 (1), S. 118–122. DOI: 10.1007/s12599-008-0005-4.

Cardelli, Luca; Wegner, Peter (1985): On understanding types, data abstraction, and polymorphism. In: *ACM Comput. Surv.* 17 (4), S. 471–523. DOI: 10.1145/6041.6042.

Ceravolo, Paolo; Junior, Sylvio Barbon; Damiani, Ernesto; van der Aalst, Wil (2024): Tuning Machine Learning to Address Process Mining Requirements. In: *IEEE Access* 12, S. 24583–24595. DOI: 10.1109/ACCESS.2024.3361650.

Chen, Ming-Syan; Han, Jiawei; Yu, P. S. (1996): Data mining: an overview from a database perspective. In: *IEEE Trans. Knowl. Data Eng.* 8 (6), S. 866–883. DOI: 10.1109/69.553155.

Chernysheva, K.; Karpuzova, N.; Afanasyeva, S.; Korolkova, A. (2021): Using the Loginom analytical platform to process long-term field experience data. In: *J. Phys.: Conf. Ser.* 2131 (3), S. 32078. DOI: 10.1088/1742-6596/2131/3/032078.

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

Cooper, Harris M. (1988): Organizing knowledge syntheses: A taxonomy of literature reviews. In: *Knowledge in Society* 1 (1), S. 104–126. DOI: 10.1007/BF03177550.

Cullen, Theresa; Garcia, Jean E. (2021): Data Mining, Data Analytics, and Bioinformatics. In: Samuel O. Okpaku (Hg.): Innovations in Global Mental Health. Cham: Springer International Publishing, S. 455–488.

Cupek, Rafal; Drewniak, Marek; Steclik, Tomasz (2021): Data Preprocessing, Aggregation and Clustering for Agile Manufacturing Based on Automated Guided Vehicles. In: Maciej Paszynski, Dieter Kranzlmüller, Valeria V. Krzhizhanovskaya, Jack J. Dongarra und Peter

M.A. Sloot (Hg.): Computational Science – ICCS 2021, Bd. 12745. Cham: Springer International Publishing (Lecture Notes in Computer Science), S. 458–470.

Dehmer, Matthias (2005): Data Mining-Konzepte und graphentheoretische Methoden zur Analyse hypertextueller Daten. In: *JLCL* 20 (1), S. 113–141. DOI: 10.21248/jlcl.20.2005.71.

Demirbaga, Ümit; Aujla, Gagangeet Singh; Jindal, Anish; Kalyon, Oğuzhan (2024): Big Data Analytics. Cham: Springer Nature Switzerland.

Dietrich, Manfred (2021): Digitales Shopfloor Management in SAP-Systemumgebungen. Wiesbaden: Springer Fachmedien Wiesbaden.

D'Onofrio, Sara; Meier, Andreas (2021): Big Data Analytics. Wiesbaden: Springer Fachmedien Wiesbaden.

Dunckern, Christian (2017): Automobilproduktion im Zeitalter der Digitalisierung. In: Alexandra Hildebrandt und Werner Landhäußer (Hg.): CSR und Digitalisierung. Berlin, Heidelberg: Springer Berlin Heidelberg (Management-Reihe Corporate Social Responsibility), S. 101–113.

Eichstädt, Timm; Spieker, Stefan (2024): 52 Stunden Informatik. Wiesbaden: Springer Fachmedien Wiesbaden.

El Morr, Christo; Jammal, Manar; Ali-Hassan, Hossam; EI-Hallak, Walid (2022): Machine Learning for Practical Decision Making. Cham: Springer International Publishing (334).

Fan, Cheng; Chen, Meiling; Wang, Xinghua; Wang, Jiayuan; Huang, Bufu (2021): A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. In: *Front. Energy Res.* 9, Artikel 652801. DOI: 10.3389/fenrg.2021.652801.

Fasel, Daniel; Meier, Andreas (2016): Big Data. Wiesbaden: Springer Fachmedien Wiesbaden.

Fayyad, Usama M.; Piatetsky-Shapiro, Gregory; Smyth, Padhraic; Uthurusamy, Ramasamy (Hg.) (1996): Advances in knowledge discovery and data mining. 5. print. Menlo Park, Calif.: AAAI Press.

Franz, Dietrich; Mattes, Rüdiger (1991): Elektronische Datenverarbeitung. EDV-System, Computertypen, Programmiersprachen, Software, Datennetze, Datenschutz. Wiesbaden: Gabler Verlag (Springer eBook Collection Business and Economics).

Frenay, Benoit; Hofmann, Daniela; Schulz, Alexander; Biehl, Michael; Hammer, Barbara (2014): Valid interpretation of feature relevance for linear data mappings. In: 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). Orlando, FL, USA, 09.12.2014 - 12.12.2014: IEEE, S. 149–156.

Fuchs, Martina; Westermeyer, Johannes (2023): Automobilindustrie. In: Elmar Kulke (Hg.): Wirtschaftsgeographie Deutschlands. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 215–225.

Gambardella, Gennaro; Di Bernardo, Diego (2019): A Tool for Visualization and Analysis of Single-Cell RNA-Seq Data Based on Text Mining. In: *Frontiers in genetics* 10, S. 734. DOI: 10.3389/fgene.2019.00734.

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

Giobergia, Flavio; Baralis, Elena; Camuglia, Maria; Cerquitelli, Tania; Mellia, Marco; Neri, Alessandra et al. (2018): Mining Sensor Data for Predictive Maintenance in the Automotive Industry. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). Turin, Italy, 01.10.2018 - 03.10.2018: IEEE, S. 351–360.

Gluchowski, Peter (2024): Datenmanagement und Datenanalyse. Wiesbaden: Springer Fachmedien Wiesbaden.

Gluchowski, Peter; Schieder, Christian; Chamoni, Peter (2021): Methoden des Data Mining für Big Data Analytics. In: Sara D'Onofrio und Andreas Meier (Hg.): Big Data Analytics. Wiesbaden: Springer Fachmedien Wiesbaden (Edition HMD), S. 25–48.

Goll, Joachim; Stamm, Tobias (2024): C als erste Programmiersprache. Wiesbaden: Springer Fachmedien Wiesbaden.

Güting, Ralf Hartmut; Dieker, Stefan (2018): Programmiersprachliche Konzepte für Datenstrukturen. In: Ralf Hartmut Güting und Stefan Dieker (Hg.): Datenstrukturen und Algorithmen. Wiesbaden: Springer Fachmedien Wiesbaden, S. 39–61.

Han, Jiawei; Kamber, Micheline; Pei, Jian (2011): Data mining. Concepts and techniques. 3rd ed (Online-Ausg.). Amsterdam, Boston, Heidelberg: Elsevier Morgan Kaufmann (The Morgan Kaufmann Series in Data Management Systems Series). Online verfügbar unter https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=729031.

He, Yulin; Ye, Xuan; Huang, Defa; Fournier-Viger, Philippe; Huang, Joshua Zhexue (2023): A Hybrid Method to Measure Distribution Consistency of Mixed-Attribute Datasets. In: *IEEE Trans. Artif. Intell.* 4 (1), S. 182–196. DOI: 10.1109/TAI.2022.3151724.

Heinrich, Bernd; Hristova, Diana; Klier, Mathias; Schiller, Alexander; Szubartowicz, Michael (2018): Requirements for Data Quality Metrics. In: *J. Data and Information Quality* 9 (2), S. 1–32. DOI: 10.1145/3148238.

Hemavathi, D.; Kumar, V. Ramesh; Regin, R.; Rajest, S. Suman; Phasinam, Khongdet; Singh, Saurabh (2021): Technical Support for Detection and Prediction of Rainfall. In: 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC). 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC). Trichy, India, 07.10.2021 - 09.10.2021: IEEE, S. 1629–1634.

Hering, Ekbert (2023): Sensorsysteme. In: Ekbert Hering und Gert Schönfelder (Hg.): Sensoren in Wissenschaft und Technik. Wiesbaden: Springer Fachmedien Wiesbaden, S. 1–3.

Hildebrand, Knut; Gebauer, Marcus; Mielke, Michael (2021): Daten- und Informationsqualität. Wiesbaden: Springer Fachmedien Wiesbaden.

Hossain, Al Amin; Lee, Seung-Jin; Huh, Eui-Nam (2013): Shear-Based Spatial Transformation to Protect Proximity Attack in Outsourced Database. In: 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications. 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Melbourne, Australia, 16.07.2013 - 18.07.2013: IEEE, S. 1633–1638.

Hossen, Md. Sharif (2020): Data Preprocess. In: Uma N. Dulhare, Khaleel Ahmad und Khairol Amali Bin Ahmad (Hg.): Machine Learning and Big Data: Wiley, S. 71–103.

Huber, Walter (2024): Industrie 4.0 kompakt – Wie Technologien unsere Wirtschaft und unsere Unternehmen verändern. Wiesbaden: Springer Fachmedien Wiesbaden.

Jassim, Mustafa Abdalrassual; Abdulwahid, Sarah N. (2021): Data Mining preparation: Process, Techniques and Major Issues in Data Analysis. In: *IOP Conf. Ser.: Mater. Sci. Eng.* 1090 (1), S. 12053. DOI: 10.1088/1757-899X/1090/1/012053.

Jiang, Mon-Fong; Tseng, Shian-Shyong; Liao, Shan-Yi (1999): Data types generalization for data mining algorithms. In: 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028). IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics. Tokyo, Japan, 12-15 Oct. 1999: IEEE, S. 928–933.

Kahraman, Cengiz; Haktanır, Elif (2023): Intelligent Systems in Digital Transformation. Cham: Springer International Publishing (549).

Kim, Byung-Woo (2024): Data Preprocessing Methods - Strategies and Best Practices: Investigating strategies and best practices for preprocessing data, including cleaning, transformation, and feature engineering. In: *Australian Journal of Machine Learning Research & Applications,* (4), S. 208–214. Online verfügbar unter https://sydneyacademics.com/index.php/ajmlra/article/view/97, zuletzt geprüft am 16.12.24.

Kiran, Ajmeera; Vasumathi, D. (2020): Data Mining: Min–Max Normalization Based Data Perturbation Technique for Privacy Preservation. In: K. Srujan Raju, A. Govardhan, B. Padmaja Rani, R. Sridevi und M. Ramakrishna Murty (Hg.): Proceedings of the Third International Conference on Computational Intelligence and Informatics, Bd. 1090. Singapore: Springer Singapore (Advances in Intelligent Systems and Computing), S. 723–734.

Koteeswaran, S.; Visu, P.; Silambarasan, K.; Vimal Karthick, R. (2014): HADOOP+Big Data: Analytics Using Series Queue with Blocking Model. In: *RJASET* 8 (3), S. 341–345. DOI: 10.19026/rjaset.8.978.

Kropik, Markus (2021): Produktionsleitsysteme für die Automobilindustrie. Berlin, Heidelberg: Springer Berlin Heidelberg.

Kumar, G. Sathish; Premalatha, K. (2021): Securing private information by data perturbation using statistical transformation with three dimensional shearing. In: *Applied Soft Computing* 112, S. 107819. DOI: 10.1016/j.asoc.2021.107819.

Kumari, Rani; Saini, Kavita (2021): Advanced Automobile Manufacturing. An Industry 4.0. In: *8th International Conference on Computing for Sustainable Global Development (INDIA-Com)*, S. 899–904.

Kurpiela, Steffen; Teuteberg, Frank (2023): Strategische Planung in der Automobilindustrie in Zeiten von Produkt-Service-Systemen: Veränderungen, Planungsansätze und Herangehensweisen. In: *HMD* 60 (6), S. 1237–1252. DOI: 10.1365/s40702-023-01013-8.

Landes, Miriam; Steiner, Eberhard (2013): Psychologie der Wirtschaft. Wiesbaden: Springer Fachmedien Wiesbaden.

Lee, Jongsuk; Chua, Ping Chong; Chen, Lequn; Ng, Poh Huat Nicholas; Kim, Yerim; Wu, Qiong et al. (2023): Key Enabling Technologies for Smart Factory in Automotive Industry: Status and Applications. In: *Int. J. Precis. Eng. Manuf.-Smart Tech.* 1 (1), S. 93–105. DOI: 10.57062/ijpem-st.2022.0017.

Loebbecke, Claudia; Picot, Arnold (2015): Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. In: *The Journal of Strategic Information Systems* 24 (3), S. 149–157. DOI: 10.1016/j.jsis.2015.08.002.

Mane, Pravin (2023): Application of Data Mining Techniques for Measuring and Predicting Employee Performances in Automotive Industry. In: *American Journal Of Economics And Business Management* (6), Artikel 1, S. 10–18. Online verfügbar unter https://www.researchgate.net/publication/371449432_Application_of_Data_Mining_Techniques_for_Measuring_and_Predicting_Employee_Performances_in_Automotive_Industry, zuletzt geprüft am 05.01.2025.

Mehmood, Mavra; Javed, Talha; Nebhen, Jamel; Abbas, Sidra; Abid, Rabia; Reddy Bojja, Giridhar; Rizwan, Muhammad (2022): A Hybrid Approach for Network Intrusion Detection. In: *Computers, Materials & Continua* 70 (1), S. 91–107. DOI: 10.32604/cmc.2022.019127.

Mertens, Peter; Buxmann, Peter; Hess, Thomas; Hinz, Oliver; Muntermann, Jan; Schumann, Matthias (2023): Grundzüge der Wirtschaftsinformatik. Berlin, Heidelberg: Springer Berlin Heidelberg.

Mielebacher, Jörg (2024): Datenbanken für Nichtinformatiker. Wiesbaden: Springer Fachmedien Wiesbaden.

Missonnier, Ruken (2024): Manufacturing Data for Polynomial Regression. Enhancing Manufacturing Efficiency through Statistical Modeling. Online verfügbar unter https://www.kaggle.com/datasets/rukenmissonnier/manufacturing-data-for-polynomial-regression/data, zuletzt geprüft am 11.02.2025.

Mondal, Kartick Chandra; Nandy, Biswadeep Deb; Baidya, Arunima (2018): A Factual Analysis of Improved Python Implementation of Apriori Algorithm. In: Jyotsna Kumar Mandal, Somnath Mukhopadhyay, Paramartha Dutta und Kousik Dasgupta (Hg.): Methodologies and Application Issues of Contemporary Computing Framework. Singapore: Springer Singapore, S. 139–151.

Moturi, Sireesha; Rao, S. N. Tirumala; Vemuru, Srikanth (2021): Grey wolf assisted dragonfly-based weighted rule generation for predicting heart disease and breast cancer. In: *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 91, S. 101936. DOI: 10.1016/j.compmedimag.2021.101936.

Mulenga, Mwenge; Phiri, Musa; Simukonda, Luckson; Alaba, Fadele A. (2023): A Multistage Hybrid Deep Learning Model for Enhanced Solar Tracking. In: *IEEE Access* 11, S. 129449–129466. DOI: 10.1109/ACCESS.2023.3333895.

Mustafa, Ahmed Bani (2019): Enhancing Learning from Imbalanced Classes via Data Preprocessing: A Data-Driven Application in Metabolomics Data Mining. In: *The ISC Int'l Journal of: ISeCure* (11), S. 79–89.

North, Klaus (2021): Wissensorientierte Unternehmensführung. Wiesbaden: Springer Fachmedien Wiesbaden.

Olson, David L.; Wu, Desheng Dash; Luo, Cuicui; Nabavi, Majid (2024): Business Analytics with R and Python. Singapore: Springer Nature Singapore.

Omar, Nashuha; Supriyanto, Eko; Wahab, Asnida Abdul; Al-Ashwal, Rania Hussien Ahmed; Nazirun, Nor Nisha Nadhira (2022): Application of K-Means Algorithm on Normalized and Standardized Data for Type 2 Diabetes Subclusters. In: 2022 International Conference on Healthcare Engineering (ICHE). 2022 International Conference on Healthcare Engineering (ICHE). Johor, Malaysia, 23.09.2022 - 25.09.2022: IEEE, S. 1–6.

Pasha, Akram; Latha, P. H. (2019): Efficient Sensory Data Transformation: A Big Data Approach. In: Radek Silhavy, Petr Silhavy und Zdenka Prokopova (Hg.): Intelligent Systems Applications in Software Engineering, Bd. 1046. Cham: Springer International Publishing (Advances in Intelligent Systems and Computing), S. 69–82.

Patro, S. Gopal Krishna; Sahu, Kishore Kumar (2015): Normalization: A Preprocessing Stage.

Petrova, Pepa; Jotsov, Vladimir; Sgurev, Vassil (2018): Puzzle Methods for Automatic Selection of Data Cleansing Techniques. In: 2018 International Conference on Intelligent Systems (IS). 2018 International Conference on Intelligent Systems (IS). Funchal - Madeira, Portugal, 25.09.2018 - 27.09.2018: IEEE, S. 820–826.

Pfister, Hans-Rüdiger; Jungermann, Helmut; Fischer, Katrin (2017): Die Psychologie der Entscheidung. Berlin, Heidelberg: Springer Berlin Heidelberg.

Possner, Lucas; Bahr, Lukas; Roehl, Leonard; Wehner, Christoph; Groeger, Sophie (2024): Anwendung von Causal-Discovery-Algorithmen zur Root-Cause-Analyse in der Fahrzeugmontage.

Rahm, Erhard; Do, Hong Hai (2000): Data Cleaning: Problems and Current Approaches. In: *IEE Data Eng. Bull*.

Raja, P. S.; Thangavel, K. (2016): Soft Clustering Based Missing Value Imputation. In: S. Subramanian, R. Nadarajan, Shrisha Rao und Shina Sheen (Hg.): Digital Connectivity – Social Impact, Bd. 679. Singapore: Springer Nature Singapore (Communications in Computer and Information Science), S. 119–133.

Runkler, Thomas A. (2015): Data mining. Modelle und Algorithmen intelligenter Datenanalyse. 2., aktualisierte Auflage. Wiesbaden: Springer Fachmedien (Lehrbuch).

Saranya, G.; Pravin, A. (2023): A novel feature selection approach with integrated feature sensitivity and feature correlation for improved prediction of heart disease. In: *J Ambient Intell Human Comput* 14 (9), S. 12005–12019. DOI: 10.1007/s12652-022-03750-y.

Sathya Durga, V.; Jeyaprakash, Thangakumar (2019): Data Transformation Techniques for Academic Datasets. In: *IJEAT* 9 (1), S. 2214–2218. DOI: 10.35940/ijeat.A9711.109119.

Schmidt, Horst (1993): Die Datenverarbeitung im Versicherungsbetrieb. Wiesbaden, s.l.: Gabler Verlag.

Schuh, Günther; Graf, Leonie; Zeller, Paul; Scholz, Paul; Studerus, Bastian (2019): Eine Branche im Umbruch - Den technologischen Wandel in der Automobilindustrie gestalten.

Schütze, Andreas; Helwig, Nikolai (2017): Sensorik und Messtechnik für die Industrie 4.0. In: *tm - Technisches Messen* 84 (5), S. 310–319. DOI: 10.1515/teme-2016-0047.

Senderovich, Arik; Shleyfman, Alexander; Weidlich, Matthias; Gal, Avigdor; Mandelbaum, Avishai (2018): To aggregate or to eliminate? Optimal model simplification for improved process performance prediction. In: *Information Systems* 78, S. 96–111. DOI: 10.1016/j.is.2018.04.003.

Siek, Michael; Mukti, Ryan Malik Gunadharma (2020): Process Mining with Applications to Automotive Industry. In: *IOP Conf. Ser.: Mater. Sci. Eng.* 924 (1), S. 12033. DOI: 10.1088/1757-899X/924/1/012033.

Siraj, Maheyzah Md; Rahmat, Nurul Adibah; Din, Mazura Mat (2019): A Survey on Privacy Preserving Data Mining Approaches and Techniques. In: Proceedings of the 2019 8th International Conference on Software and Computer Applications. ICSCA '19: 2019 8th International Conference on Software and Computer Applications. Penang Malaysia, 19 02 2019 21 02 2019. New York, NY, USA: ACM, S. 65–69.

Steiner, René (2014): Grundkurs Relationale Datenbanken. Wiesbaden: Springer Fachmedien Wiesbaden.

Steyer, Ralph (2024): Programmierung in Python. Wiesbaden: Springer Fachmedien Wiesbaden.

Strong, Diane M.; Lee, Yang W.; Wang, Richard Y. (1997): Data quality in context. In: *Commun. ACM* 40 (5), S. 103–110. DOI: 10.1145/253769.253804.

Tidake, V. S.; Sane, S. S. (2018): Evaluation of Multi-label Classifiers in Various Domains Using Decision Tree. In: Subhash Bhalla, Vikrant Bhateja, Anjali A. Chandavale, Anil S. Hiwale und Suresh Chandra Satapathy (Hg.): Intelligent Computing and Information and Communication, Bd. 673. Singapore: Springer Singapore (Advances in Intelligent Systems and Computing), S. 117–127.

Tränkler, Hans-Rolf; Aronshtam, Yulia (2014): Signalverarbeitung bei Einzelsensoren. In: Hans-Rolf Tränkler und Leo Reindl (Hg.): Sensortechnik. Berlin, Heidelberg: Springer Berlin Heidelberg (VDI-Buch), S. 1175–1204.

Ungermann, Florian; Kuhnle, Andreas; Stricker, Nicole; Lanza, Gisela (2019): Entscheidungsunterstützungs-systeme in der Produktion. In: *Zeitschrift für wirtschaftlichen Fabrikbetrieb* 114 (1-2), S. 34–38. DOI: 10.3139/104.112034.

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

Vogel-Heuser, Birgit; Hompel, Michael ten; Bauernhansl, Thomas (2024): Handbuch Industrie 4.0. Berlin, Heidelberg: Springer Berlin Heidelberg.

vom Brocke, Jan; Simons, Alexander; Niehaves, Bjoern; Reimer, Kai (2009): Reconstruction The Giant. On The Importance Of Rigour In Documenting The Literature Search Process. In: *ECIS 2009 Proceedings. 161*.

Weiss, Mark Allen (2013): Data structures and algorithm analysis in C++. Fourth edition. Boston, London: Pearson.

Wildemann, Horst (2013): Entwicklungslinien der Produktionssysteme in der Automobilindustrie. In: Ingrid Göpfert, David Braun und Matthias Schulz (Hg.): Automobillogistik. Wiesbaden: Springer Fachmedien Wiesbaden, S. 135–158.

Winkelhake, Uwe (2024): Die digitale Transformation der Automobilindustrie. Berlin, Heidelberg: Springer Berlin Heidelberg.

Yang, Chao-Lung; Yang, Chen-Yi; Chen, Zhi-Xuan; Lo, Nai-Wei (2019): Multivariate Time Series Data Transformation for Convolutional Neural Network. In: 2019 IEEE/SICE International Symposium on System Integration (SII). 2019 IEEE/SICE International Symposium on System Integration (SII). Paris, France, 14.01.2019 - 16.01.2019: IEEE, S. 188–192.

Zhang, S. Z.; Yu, Y. R.; Shen, M. Z. (2021): Study on Preprocessing Method of TCM Prescription Data in Data Mining. In: *J. Phys.: Conf. Ser.* 2025 (1), S. 12027. DOI: 10.1088/1742-6596/2025/1/012027.

Zhen, Chenggang; Zhang, Yingxin (2019): A Survey of Data Preprocessing in Data Mining. In: *International Core Journal of Engineering* 5 (9), S. 133–137.

Zheng, Jie; Harris, Marcelline R.; Masci, Anna Maria; Lin, Yu; Hero, Alfred; Smith, Barry; He, Yongqun (2016): The Ontology of Biological and Clinical Statistics (OBCS) for standardized

and reproducible statistical analysis. In: *Journal of biomedical semantics* 7 (1), S. 53. DOI: 10.1186/s13326-016-0100-2.

Zwiener, Isabella; Frisch, Barbara; Binder, Harald (2014): Transforming RNA-Seq data to improve the performance of prognostic gene signatures. In: *PloS one* 9 (1), e85150. DOI: 10.1371/journal.pone.0085150.

Anhang A: Systematische Literaturrecherche

A.1 Übersicht der systematischen Literaturrecherche

Index	Titel	Autoren	Jahr
[1]	What Does Data Analytics Offer for Extracting Knowledge from Middle-of-Life Product Data?	(Abou Eddahab und Horvath)	2018
[2]	Predicting lung cancer survival based on clinical data using machine learning: A review	(Altuhaifa et al.)	2023
[3]	Intelligent ensemble of voting based solid fuel classification model for energy harvesting from agricultural residues	(Al-Wesabi et al.)	2022
[4]	Ridge Regularized Imputed Scaled Clipping Normalization Based Pre-processing for Marine Weather Forecasting	(Anbarasi und Radha)	2023
[5]	Imbalanced data classification using improved synthetic minority over-sampling technique	(Anusha et al.)	2023
[6]	Benchmarking of cell type deconvolution pipelines for transcriptomics data	(Avila Cobos et al.)	2020
[7]	DATM: A Novel Data Agnostic Topic Modeling Technique With Improved Effectiveness for Both Short and Long Text	(Bewong et al.)	2023
[8]	Tuning Machine Learning to Address Process Mining Requirements	(Ceravolo et al.)	2024
[9]	Using the Loginom analytical platform to process long-term field experience data	(Chernysheva et al.)	2021
[10]	Data Mining, Data Analytics, and Bioinformatics: Leveraging Big Data to Identify the Most Vulnerable and to Reduce Health Disparities	(Cullen und Garcia)	2021
[11]	Valid interpretation of feature relevance for linear data mappings	(Frenay et al.)	2015
[12]	A tool for visualization and analysis of single-cell RNA-seq data based on text mining	(Gambardella und Di Bernardo)	2019
[13]	A Hybrid Method to Measure Distribution Consistency of Mixed-Attribute Datasets	(He et al.)	2023
[14]	Technical Support for Detection and Prediction of Rainfall	(Hemavathi et al.)	2021
[15]	Data Preprocess	(Hossen)	2020

Index	Titel	Autoren	Jahr
[16]	Data mining: Min–max normalization-based data perturbation technique for privacy preservation	(Kiran und Vasumathi)	2020
[17]	HADOOP+big data: Analytics using series queue with blocking model	(Koteeswaran et al.)	2014
[18]	Securing private information by data perturbation using statistical transformation with three dimensional shearing [Formula presented]	(Kumar und Premalatha)	2021
[19]	A hybrid approach for network intrusion detection	(Mehmood et al.)	2021
[20]	A factual analysis of improved python implementation of apriori algorithm	(Mondal et al.)	2018
[21]	Grey wolf assisted dragonfly-based weighted rule generation for predicting heart disease and breast cancer	(Moturi et al.)	2021
[22]	A Multistage Hybrid Deep Learning Model for Enhanced Solar Tracking	(Mulenga et al.)	2023
[23]	Enhancing Learning from Imbalanced Classes via Data Preprocessing: A Data-Driven Application in Metabolomics Data Mining☆	(Mustafa)	2019
[24]	Application of K-Means Algorithm on Normalized and Standardized Data for Type 2 Diabetes Subclusters	(Omar et al.)	2022
[25]	Efficient Sensory Data Transformation: A Big Data Approach	(Pasha und Latha)	2019
[26]	Data Mining in the Mixed Methods: Application to the Study of the Psychological Profiles of Athletes	(Petrova et al.)	2018
[27]	Soft clustering based missing value imputation	(Raja und Thangavel)	2016
[28]	A novel feature selection approach with integrated feature sensitivity and feature correlation for improved prediction of heart disease	(Saranya und Pravin)	2023
[29]	Data transformation techniques for academic datasets	(Sathya Durga und Jeyaprakash)	2019
[30]	Evaluation of multi-label classifiers in various domains using decision tree	(Tidake und Sane)	2018
[31]	Study on preprocessing method of TCM prescription data in data mining	(Zhang et al.)	2021
[32]	The Ontology of Biological and Clinical Statistics (OBCS) for standardized and reproducible statistical analysis	(Zheng et al.)	2016

A.2 Prüfung der Literatur auf die Verfahren

Index	Coding	Generalization	Aggregation	Binning	Discretization	Normalization	Clustering	Regression	Unbekan
[1]			×						×
[2]	×			×		×			
[3]	×					×		×	
[4]						×		×	×
[5]		×	×			×	×		×
[6]			×			×	×	×	×
[7]			×						×
[8]	×	×	×				×	×	
[9]			×						×
[10]			×						×
[11]			×		×	×		×	
[12]						×	×		×
[13]	×					×	×	×	×
[14]		×	×			×		×	
[15]		×	×	×	×	×	×	×	
[16]						×			
[17]							×		×
[18]						×			×
[19]	×					×			
[20]	×					×	×	×	
[21]						×			
[22]			×			×			×
[23]						×			×
[24]						×	×		
[25]			×			×			×
[26]						×			×
[27]			×				×		
[28]						×		×	
[29]			×			×			×
[30]	×								
[31]						×			
[32]						×			

A.3 Unbekannte Verfahren aus den Quellen

Index	Unbekanntes Verfahren
[1]	Mathematic Transformations: Logarithm, Square Root, Squaring, Reciprocal
[4]	Scaled-Clipping-Normalized-Data-Transformation
[5]	Smoothing
[6]	Column- and Global-Min-Max-Normalization, Row Wise, Column- and Global-Z-Score-Normalizatiom, Quantile Normalization, Log Normalization,
[7]	Self- und Auxiliary Aggregation (Textaggregation)
[9]	Program (Line filter, Cross-table, Cross-Chart, Sliding Window)
[10]	Datatyp Transformation
[12]	Euclidian Normalization
[13]	Euclidian-Distance-For-Similarity
[17]	Map Reducer
[18]	Axis-Transformation
[22]	Sigmoid Normalization, Sine- and Cosine-Transformation, Gramian-Angular-Field-Encoding
[23]	Log Transformation
[25]	Data-Structure-Transformation
[26]	Linearization, Box-Cox-Transformation
[29]	Cube-Root-Normalization, Decimal-Scaling-Normalization, Normalization-Using-Norm, Square-Root-Normalization

A.4 Rückwärtssuche zur Datentransformation aus den Publikationen der Literaturrecherche

Index aus Anhang 1	Quelle der Publikation	Index für Publikationsquelle	Datentransformationsverfahren aus der zitierten Quelle
[1]	Rahm und Do (2000).Data Cleaning: Problems and Current Approaches.	[1a]	Aggregation, Standardization, Generalization, Clustering,
	Hossain et al. (2013). Shear-based Spatial Transformation to Protect Proximity Attack in Outsourced Database	[1b]	Spatial-Transformation, Min-Max-Normalization, Clustering
[2]	Chen et al. (1996). Data mining: an overview from a database perspective	[2a]	Aggregation, Generalization, Normalization, Clustering, Regression
	Siraj et al. (2019). A survey on privacy preserving data mining approaches and techniques	[2b]	Normalization, Clustering, Aggregation, Regression
[6]	Zwiener et al. (2014). Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures	[6a]	Regression, Log Transformation, Variance-Stabilizing-Transformation, Box-Cox-Transformation, Standardization, Normalization, Standardized Logs, Standardized-Variance-Stability-Transformation, Standardized-Box-Cox-Transformation, Rank Transformation, Blom Transformation
[8]	Senderovich et al. (2018) 'To aggregate or to eliminate? Optimal model simplification for improved process performance prediction	[8a]	Aggregation, Logical Aggregation
[13]	Aggarwal (2017). Data Mining	[13a]	Coding, Aggregation, Generalization, Binning, Discretization, Normalization, Clustering, Regression,
[15]	Han et al. (2011). Data Mining: Concepts and Techniques	[15a]	Coding, Aggregation, Generalization, Binning, Discretization, Normalization, Clustering, Regression

Index aus Anhang 1	Quelle der Publikation	Index für Publikationsquelle	Datentransformationsverfahren aus der zitierten Quelle
	García et al. (2015). Data Preprocessing in Data Mining	[15b]	Coding, Aggregation, Generalization, Binning, Discretization, Normalization, Clustering, Regression, Mathematic Transformations: Linear- and Quadratic-transformation, Non-Polynomial-Approximations-Of-Transformations, Polynomial-Approximations-Of-Transformations, Rank Transformation, Box-Cox-Transformations
[22]	Yang et al. (2019) 'Multivariate time series data transformation for convolutional neural network	[22a]	Coding
[24]	García et al. (2015). Data Preprocessing in Data Mining	[24a]	s. bei [15] in gleicher Tabelle
[29]	Patro und Sahu (2015). Normalization: A Preprocessing Stage	[29a]	Min-Max-Normalization, Z-Score-Normalization, Decimal-Scaling-Normalization, Integer-Scaling-Normalization
[31]	Alasadi und Bhaya (2017). Review of Data Preprocessing techniques in Data Minng	[31a]	Binning (Bin Means, Equal-Depth-Bins, Bin Boundaries), Regression, Clustering, Data Cube Aggregation, Mathematic Aggregation, Generalization, Normalization, Discretization (Binning, Histogram, Entropy-Based, Clustering)
	Zhen und Zhang (2019). A Survey of data Preprocessing in Data Mining	[31b]	Aggregation (Dimension Reduction)

Anhang B: Algorithmen der exemplarischen Validierung aus Python

B.1 Dicretization der Qualität

Algorithmus 1: Discretization der Qualität aus dem ersten Szenario

```
# Einfügen der Pandas Bib

import pandas as pd

# Definieren des Datensatzes der diskretisiert werden soll mit Pfad

Datensatz_VorDiscr = r"C:\Users\MSI Thin\Desktop\Mappe2.csv"

# Datensatz wird gelesen und strukturiert

Datensatz = pd.read_csv(Datensatz_VorDiscr, delimiter=";", decimal=",")

# als Dataframe wird nur die Spalte Quality Rating betrachtet

Datensatz = Datensatz[["Quality Rating"]]

# Dateityp wird in float umgewandelt

Datensatz["Quality Rating"] = Datensatz["Quality Rating"].astype(float)

# Grenzen der Intervalle

Schwellenwerte = [0, 76, 91, 100]

# Skala der Bewertung

Bewertung = ["Niedrig", "Mittel", "Hoch"]

# Die Quality Ratingwerte werden kategorisiert mithilfe der Schwellenwerte und der vorgegebenen Bewertung

Datensatz["Quality Category"] = pd.cut(Datensatz["Quality Rating"], bins=Schwellenwerte, labels=Bewertung, right=True)

# Die Ersten 20 Zeilen wiedergeben, zum Check

print(Datensatz.head(20))

# Datei wird nach Diskretisierung gespeichert

Datensatz.to_csv("Datensatz_NachDiscr.csv", index=False, sep=";")
```

B.2 Discretization der Prozessparameter und der Qualität

Algorithmus 2: Discretization der Prozessparameter und der Qualität aus dem ersten Szenario

```
# Einfügen der Pandas Bib

import pandas as pd

# Definieren des Datensatzes der diskretisiert werden soll mit Pfad

Datensatz_VorDiscr = r"C:\Users\MSI Thin\Desktop\test.csv"

# Datensatz wird gelesen und strukturiert

Datensatz = pd.read_csv(Datensatz_VorDiscr, delimiter=";", decimal=",")

# Die Spalten Quality Rating, Temperatur und Druck werden beachtet

Datensatz = Datensatz[["Quality Rating", "Temperature (°C)", "Pressure (kPa)"]]

# Spalten werde in numerische Werte umgewandelt, falls es nicht möglich ist, wird NaN eingesetzt

Datensatz["Quality Rating"] = pd.to_numeric(Datensatz["Quality Rating"], errors="coerce")

Datensatz["Temperature (°C)"] = pd.to_numeric(Datensatz["Temperature (°C)"], errors="coerce")

Datensatz["Pressure (kPa)"] = pd.to_numeric(Datensatz["Pressure (kPa)"], errors="coerce")

# Entfernen der Zeilen NaN

Datensatz = Datensatz.dropna()

# Schwellenwerte der Qualität

Schwellenwerte_Qualitaet = [0, 76, 91, 100]

# Bewertungsskala des Quality Rating

Labels_Qualitaet = ["Niedrig", "Mittel", "Hoch"]

# Der Prozess der Discretization wird durchgeführt für Quality Rating

Datensatz["Quality Category"] = pd.cut(Datensatz["Quality Rating"], bins=Schwellenwerte_Qualitaet,
labels=Labels_Qualitaet, right=True)

# Discretization der Temperatur

Datensatz["Temperature Category"] = pd.cut(Datensatz["Temperature (°C)"],

Schwellenwerte_Temperatur=[100, 166.66, 233.33, 300],
```

Algorithmus 2: Discretization der Prozessparameter und der Qualität aus dem ersten Szenario

```
Labels_Temperatur=["Niedrig", "Mittel", "Hoch"],

include_lowest=True)

# Discretization des Drucks

Datensatz["Pressure Category"] = pd.cut(Datensatz["Pressure (kPa)"],

Schwellenwerte_Druck=[5, 11.66, 18.33, 25],

Labels_Druck=["Niedrig", "Mittel", "Hoch"],

include_lowest=True)

# 20 Zeilen werden angezeigt

print(Datensatz.head(20))

# Neuer Datensatz wird auf dem Desktop gespeichert

Datensatz_NachDiscr = r"C:\Users\MSI Thin\Desktop\diskretisierter_datensatz.csv"

#Neuer Datensatz immernoch im CSV Format mit Semikolon als Trennzeichen

Datensatz.to_csv(Datensatz_NachDiscr, index=False, sep=";")
```

B.3 Assoziationsanalyse der Prozessparameter und der Qualität

Algorithmus 3: Assoziationsanalyse der diskretisierten Parameter

```
# Einfügen der Pandas Bib

import pandas as pd

# Aus der mllxtend.preproccessing wird binärer Encoder importiert

from mlxtend.preprocessing import TransactionEncoder

# Aus mlxtend.fequent_patterns wird Apriori Algorithmus und Assoziationsregeln importiert

from mlxtend.frequent_patterns import apriori, association_rules

# CSV-Datei laden und Änderung des CSV-Format

Datensatz = pd.read_csv(r"C:\Users\MSI Thin\Desktop\Test2.csv", sep=";", encoding='ISO-8859-1')

# Betrachtung der diskretisierten Splaten

Datensatz = Datensatz[["Temperature diskretisiert", "Pressure diskretisiert", "Quality Rating diskreti-
siert"]]

# Die betrachteten Spalten werden in eine Liste transformiert

Transaktionen = Datensatz.astype(str).apply(lambda row: row.tolist(), axis=1).tolist()

# TransactionEncoder auf die Liste der Transaktionen anwenden

Trafo = TransactionEncoder()

NachTrafo = Trafo.fit(Transaktionen).transform(Transaktionen)

# Umwandlung in DataFrame

Datensatz_Neu = pd.DataFrame(NachTrafo, columns=Trafo.columns_)

# Apriori Algorithmus für Assoziationsregeln mit 0.3 Support

frequent_itemsets = apriori(Datensatz_Neu, min_support=0.3, use_colnames=True)

#Regeln werden mit der Funktion association_rules und Konfidenzwert 0.7 ermittelt

Regeln=association_rules(frequent_itemsets, metric="confidence", min_threshold=0.7)

# Ergebnisse anzeigen

print(frequent_itemsets, Regeln)
```

Algorithmus 3: Assoziationsanalyse der diskretisierten Parameter

```
Regeln.to_csv("assoziationsregeln.csv", index=False)
```

B.4 Clustering ohne Normalization

Algorithmus 4: Clustering der Prozessparameter ohne Normalization

```
# Einfügen der Pandas Bib
import pandas as pd
# Einfügen der matplotlib
import matplotlib.pyplot as matplt
# K-Means wird aus Scikit-Learn-Cluster importiert
from sklearn.cluster import KMeans
# Datensatz laden.
Datensatz = pd.read_csv(r"C:\Users\MSI Thin\Desktop\Test2.csv", sep=";", encoding='ISO-8859-1')
# Spalten des Datensatzes.
Datensatz.columns = Datensatz.columns.str.strip()
Datensatz = Datensatz[['Temperature (°C)', 'Pressure (kPa)', 'Quality Rating']].dropna()
# Datentyp wird festgestellt.
Datensatz['Temperature (°C)'] = Datensatz['Temperature (°C)'].str.replace(',', '.').astype(float)
Datensatz['Pressure (kPa)'] = Datensatz['Pressure (kPa)'].str.replace(',', '.').astype(float)
Datensatz['Quality Rating'] = Datensatz['Quality Rating'].str.replace(',', '.').astype(float)
# K-Means-Algorithmus mit 3 Clustern für die Prozessparameter
K_Means_Algo = KMeans(n_clusters=3, random_state=34, max_iter=50)
Datensatz['Cluster'] = K_Means_Algo.fit_predict(Datensatz[['Temperature (°C)', 'Pressure (kPa)']])
# Visualisierung der Abbildung und Temperatur & Druck mit Farben nach Qualität
matplt.figure(figsize=(9, 5))
scatter = matplt.scatter(Datensatz['Temperature (°C)'], Datensatz['Pressure (kPa)'],
c=Datensatz['Quality Rating'], cmap='RdYlGn', edgecolors='white', alpha=0.8)
# Der Plot wird beschriftet und dargestellt
matplt.xlabel("Temperature (°C)")
```

Algorithmus 4: Clustering der Prozessparameter ohne Normalization

```
matplt.ylabel("Pressure (kPa)")

matplt.title("Datenpunkte mit Qualitätsbwertung ohne Normalization")

matplt.colorbar(label="Quality Rating")

matplt.show()
```

B.5 Clustering mit Normalization

Algorithmus 5: Clustering der Prozessparameter mit Normalization

```
# Einfügen der Pandas Bib

import pandas as pd

# Einfügen der matplotlib

import matplotlib.pyplot as matplt

# K-Means wird aus Scikit-Learn-Cluster importiert

from sklearn.cluster import KMeans

# Min-Max-Scaler wird aus Scikit-Learn-Preprocessing importiert

from sklearn.preprocessing import MinMaxScaler

# Datensatz laden.

Datensatz = pd.read_csv(r"C:\Users\MSI Thin\Desktop\Test2.csv", sep=";", encoding='ISO-8859-1')

# Spalten des Datensatzes.

Datensatz.columns = Datensatz.columns.str.strip()

Datensatz = Datensatz[['Temperature (°C)', 'Pressure (kPa)', 'Quality Rating']].dropna()

# Datentyp wird festgestellt.

Datensatz['Temperature (°C)'] = Datensatz['Temperature (°C)'].str.replace(',', '.').astype(float)

Datensatz['Pressure (kPa)'] = Datensatz['Pressure (kPa)'].str.replace(',', '.').astype(float)

Datensatz['Quality Rating'] = Datensatz['Quality Rating'].str.replace(',', '.').astype(float)

# Min-Max-Skalierung für die Prozessparameter

Skalierer = MinMaxScaler()

Datensatz[['Temperature_Norm',  'Pressure_Norm']]  =  Skalierer.fit_transform(Datensatz[['Tempera-
ture (°C)', 'Pressure (kPa)']])

# K-Means Algorithmus mit 3 Clustern

K_Means_Algo = KMeans(n_clusters=3, random_state=34, max_iter=50)

Datensatz['Cluster'] = K_Means_Algo.fit_predict(Datensatz[['Temperature_Norm', 'Pressure_Norm']])
```

Algorithmus 5: Clustering der Prozessparameter mit Normalization

```
# Visualisierung: Temperatur & Druck mit Farben nach Qualität
matplt.figure(figsize=(9, 5))
scatter = matplt.scatter(Datensatz['Temperature_Norm'], Datensatz['Pressure_Norm'],
                         c=Datensatz['Cluster'], cmap='Greys', edgecolors='black', alpha=0.8)
# Der Plot wird beschriftet und dargestellt
matplt.xlabel("Temperature (°C)")
matplt.ylabel("Pressure (kPa)")
matplt.title("Clustering mit Normalization")
matplt.show()
```

B.6 Python Bibliothek

Python Bibliothek	Quelle
Pandas	https://pandas.pydata.org/
	https://pandas.pydata.org/docs/user_guide/io.html
	https://pandas.pydata.org/docs/user_guide/dsintro.html
	https://pandas.pydata.org/docs/reference/general_functions.html
Mlxtend	https://rasbt.github.io/mlxtend/
	https://rasbt.github.io/mlxtend/user_guide/preprocessing/TransactionEncoder/
	https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/
	https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/
Scikit-Learn	https://scikit-learn.org/stable/
	https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html
	https://scikit-learn.org/stable/modules/preprocessing.html
Matplotlib	https://matplotlib.org/
	https://matplotlib.org/stable/users/explain/figure/figure_intro.html