technische universität dortmund

IT in Produktion und Logistik

Strukturierte Analyse von Verfahren zur Behandlung von Ausreißern in der Datenvorverarbeitung für das Data Mining mit anschließender Entwicklung einer Entscheidungsunterstützung zur Verfahrensauswahl

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

Name:	Jana-Maria Purrmann
Matrikelnummer:	175199
Studiengang:	Logistik

Ausgabedatum:	20.03.2023
Abgabedatum:	12.06.2023

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

Tabellenverzeichnis

Abkürzungsverzeichnis

AutoML	Automated Machine Learning
AVG-T	Schwellenwerttrimmen
CRISP-DM	Cross Industry Standard Process for Data Mining
KDD	Knowledge Discovery in Databases
kNN	k-Nearest-Neighbour
LAD	least absolute deviation
LTS	least trimmed squares
OLS	ordinary least squares
WLAD	weighted least absolute deviation

1 Einleitung

Das Volumen der gespeicherten Daten in wissenschaftlichen und wirtschaftlichen Datenbanken unterliegt einem exponentiellen Wachstum (Bensberg 2001). Dies begründet sich hauptsächlich in zwei Punkten: die Kosten für die Datenspeicherung und -verarbeitung fallen und der Alltag wird zunehmend digitalisiert. Für immer mehr Prozesse und Vorgänge besteht die Möglichkeit, Zustände und deren Änderungen in Echtzeit zu messen und digital zu speichern (Müller und Lenz 2013). Durch die immer größer werdenden Datenbestände wird für Unternehmen die Wissensentdeckung in Datenbanken, auch genannt Knowledge Discovery in Databases (KDD) relevanter. Das manuelle Untersuchen von Datenbeständen ist zeit- und kostenintensiv (Sharafi 2012). Außerdem ist durch die Größe der Datenbestände eine manuelle Verarbeitung der Daten in den meisten Fällen nicht mehr möglich. Daher ist unter Anderem der Bedarf nach computergestützten Techniken und Methoden entstanden. Mit dieser Hilfe sollen unentdeckte und nützliche Informationen aus großen Datenbeständen extrahiert werden (Sharafi 2012). Außerdem sollen noch nicht bekannte Zusammenhänge erkannt werden (Fasel und Meier 2016). In der Literatur werden die Begriffe des Data Mining und KDD nicht immer klar voneinander getrennt. Nach Fayyad, Piatetsky-Shapiro und Smith (1996) ist Data Mining aber ein Teil des KDD Prozesses. In dieser Arbeit wird das Data Mining ebenfalls als ein Teil des KDD Prozesses behandelt. Hier werden spezielle Algorithmen zur Musterextraktion verwendet. Da Data Mining Verfahren ein Teil des KDD Prozesses sind, kann die losgelöste Anwendung leicht zu der Entdeckung von bedeutungslosen oder ungültigen Mustern führen (Fayyad et al. 1996).

Ein weiterer Teil des KDD Prozesses ist die Datenvorverarbeitung. Diese stellt einen der wichtigsten Aspekte dar, da die Datenvorverarbeitung einen großen Einfluss auf die Qualität der Data Mining Ergebnisse hat (Aggarwal 2015). Die Daten müssen im Vorfeld vorverarbeitet werden, um eine effiziente und qualitativ hochwertige Analyse der Daten mit Hilfe des Data Minings erzielen zu können. Ein Bestandteil der Vorverarbeitung ist, dass aus teilweise unübersichtlichen Daten geeignete Merkmale generiert und die relevanten Daten ausgewählt werden (Runkler 2020). Die gesamte Vorverarbeitung der Daten nimmt typischerweise den größten Teil innerhalb eines Data Mining Projekts in Anspruch (Müller und Lenz 2013). Eine der wichtigsten Aufgaben der Datenvorverarbeitung ist die Erkennung und Behandlung von Fehlern sowie Ausreißern und Rauscheffekten. Bei Rauschen handelt es sich in der Regel um Mess- und Übertragungsfehler (Sharafi 2012). Ausreißer können als Werte bezeichnet werden, welche stark von den übrigen Werten abweichen (Cleve und Lämmel 2020). Es kann sich dabei sowohl um korrekt erfasste Daten handeln als auch um falsche Angaben (Cleve und Lämmel 2020). Runkler (2020) ordnet Ausreißer den Fehlern zu. Fehler in Daten lassen sich in zufällige und systematische Fehler aufteilen, wobei sich die Klasse der Ausreißer beiden Fehlerarten zuordnen lässt (Runkler 2020). Lokale Ausreißer sind allgemein als einzelne Daten definiert, die stark von den lokalen Nachbarn abweichen (Breunig et al. 2000). Im Gegensatz dazu stehen die globalen Ausreißer, welche sich dadurch auszeichnen, dass sie stark von der Verteilung der restlichen Daten abweichen (Runkler 2020). Beide Arten der Ausreißer können unter anderem mit Filtermethoden oder einer Clusteranalyse erkannt und reduziert oder gelöscht werden (Runkler 2020). Des Weiteren können die lokalen und globalen Ausreißer in Punkt-, sowie kontextuale und kollektive Ausreißer eingeteilt werden (Chandola et al. 2009). Ein Punktausreißer ist als einzelne Dateninstanz zu betrachten, welche von den anderen Dateninstanzen stark abweicht (Divya und Babu 2016). Kontextuale Ausreißer erscheinen nur in einem bestimmten Kontext als anormal und kollektive Ausreißer sind für sich genommen eine Anormalität, ihr gemeinsames Auftreten ist allerdings anormal (Chandola et al. 2009; Divya und Babu 2016). Ausreißer können allerdings nicht nur als Fehler eingeordnet werden, welche behandelt werden müssen, sondern sie können auch durchaus wertvolle Informationen beinhalten (Koufakou und Georgiopoulos 2010). Daher sollten sie vor einer Behandlung überprüft werden.

Da viele Verfahren zur Behandlung von Ausreißern zur Verfügung stehen, ist der Anwender bei der Auswahl des passenden Verfahrens mit einigen Herausforderungen konfrontiert. Durch eine Methode der Entscheidungsunterstützung kann die Verfahrenswahl zur Behandlung von Ausreißern für den Anwender erleichtert werden. Sie zeigt verschiedene Handlungsmöglichkeiten auf und trägt zu einer fundierten Entscheidung bei (Buchholz und Clausen 2009). Im Vorfeld sollten Anforderungen an die Verfahren formuliert sein. Die sich daraus ergebenden Kriterien ermöglichen eine Einordnung des zu untersuchenden Datensatzes.

Das Hauptziel dieser Arbeit besteht in der Entwicklung einer Entscheidungsunterstützung bezüglich der Verfahrenswahl zur Behandlung von Ausreißern im Kontext der Datenvorverarbeitung für das Data Mining. Dazu werden verschiedene Teilziele formuliert. Das erste Teilziel umfasst die theoretischen Grundlagen der Entstehung von Wissen sowie des Prozesses der Wissensentdeckung in Datenbanken. Insbesondere wird hier auf die Datenvorverarbeitung und Ausreißer eingegangen. Das zweite Teilziel ist die Vorstellung von Verfahren der Ausreißerbehandlung mittels einer systematischen Literaturrecherche. Mit Hilfe der beiden vorherigen Teilziele kann das dritte Teilziel erreicht werden. Dieses besteht in der Ableitung von Anforderungen an die zu entwickelnde Methode zur Entscheidungsunterstützung und der Formulierung von Kriterien. Dies soll es dem Anwender erleichtern, eine Entscheidung bezüglich des passenden Verfahrens zur Ausreißerbehandlung zu treffen. Die anschließende Entwicklung der Methode zur Entscheidungsunterstützung stellt wie bereits beschrieben das Hauptziel dar.

Um zu den zuvor genannten Zielen zu gelangen, wird in Kapitel 2 ein Überblick über den Stand der Technik gegeben. Im Zuge dessen werden unter anderem allgemeine Methoden zur Entscheidungsunterstützung vorgestellt, die Entstehung von Wissen erklärt und wichtige Begriffe aus den Themenfeldern des KDD, Data Mining und Datenvorverarbeitung besprochen und definiert. Dabei werden die Begriffe KDD und Data Mining voneinander abgegrenzt, um die Unterschiede aufzuzeigen und die Begriffe für nachfolgende Kapitel zu konkretisieren. Des Weiteren wird in diesem Abschnitt eine grundlegende Einführung in den Begriff der Ausreißer gegeben. Hier werden charakteristische Merkmale dargestellt sowie kurz auf die Ausreißerbehandlung eingegangen. In Kapitel 3 wird zunächst das Vorgehen der strukturierten Literaturrecherche beschrieben. Im Anschluss werden verschiedene Verfahren zur Ausreißerbehandlung vorgestellt. Dadurch sollen Möglichkeiten aufgezeigt werden, wie mit Ausreißern in der Datenvorverarbeitung umgegangen werden kann, um anschließend das Data Mining durchführen zu können. Die vorgestellten Verfahren werden für eine bessere Übersicht dabei in Oberbegriffe unterteilt. Außerdem werden die verschiedenen Ansichten der Autoren diskutiert. Im nachfolgenden Kapitel 4 werden mit Hilfe der vorangegangenen Literaturrecherche allgemeine Anforderungen an die Ausreißerbehandlung abgeleitet und definiert. Hierauf folgt die Bestimmung von Kriterien, um einen Vergleich unter den in Kapitel 3 vorgestellten Verfahren anstellen zu können. Nach der Formulierung der Kriterien wird auf Basis dieser eine Entscheidungsunterstützung entwickelt. Diese unterstützt den Anwender bei der Wahl des passenden Verfahrens zur Behandlung von Ausreißern. Im Anschluss werden die gewonnenen Erkenntnisse dieser Arbeit zusammengefasst und diskutiert. Insbesondere wird in diesem Teil auf die Vorteile und Schwachstellen der entwickelten Entscheidungsunterstützung eingegangen. In Kapitel 5 werden eine abschließende Zusammenfassung und ein Ausblick gegeben.

2 Wissensentdeckung in Datenbanken

In diesem Kapitel werden grundlegende Begriffe erläutert, die für den Verlauf dieser Arbeit wichtig sein werden. Zunächst wird ein Überblick über das Themenfeld von Methoden zur Entscheidungsunterstützung gegeben. Im Anschluss wird die Entstehung von Wissen in Anlehnung an die in der Literatur oft zitierte Wissenspyramide erläutert. Danach werden relevante Begriffe der Wissensentdeckung in Datenbanken definiert und erläutert. Des Weiteren werden die Begriffe des KDD und Data Mining abgegrenzt. Im vierten Teil dieses Kapitels wird detailliert auf die Datenvorverarbeitung eingegangen. Abschließend wird der Ausreißerbegriff erläutert, Charakteristika von Ausreißern beschrieben und kurz auf die allgemeinen Behandlungsmöglichkeiten von Ausreißern eingegangen. Dieses Kapitel stellt die Grundlage der darauffolgenden Kapitel dar und leitet das nachfolgende Kapitel Verfahren zur Behandlung von Ausreißern und der damit einhergehenden strukturierten Literaturrecherche ein.

2.1 Methode zur Entscheidungsunterstützung

Entscheidungen zu treffen stellt oft hohe Anforderungen an den Entscheider (Brinkmeyer und Müller 1994). Zum einen haben Entscheidungsträger in der Regel eine große Anzahl an Alternativen zur Verfügung (Spengler et al. 2017). Zum anderen ist zu beachten, dass die Ressourcen bezüglich der Zeit, dem Personal und dem Budget begrenzt sind (Spengler et al. 2017). Daher haben sich einige Methoden zur Entscheidungsfindung entwickelt, welche den Anspruch haben Entscheidungsprobleme so zu gestalten, dass diese systematisch und logisch gelöst werden können (Brinkmeyer und Müller 1994). Bei schwierigen Entscheidungen sollte zunächst eine angemessene Darstellung des Problems erfolgen. Die Methode zur Lösung des Problems sollte dabei unterstützen, das Problem zu strukturieren. Das bedeutet einzelne Elemente und Komponenten werden herausgestellt und in eine Ordnung gebracht (Brinkmeyer und Müller 1994).
Nach Bell et al. (1988) kann ein Problem in die folgenden Elemente unterteilt werden:

- Ein oder mehrere Entscheidungsträger
- Verschiedene Handlungsalternativen
- Konsequenzen der verschiedenen Handlungsalternativen
- Erstellung einer Rangfolge der Konsequenzen durch den/die Entscheidungsträger
- Darstellung der zukünftigen Zustände
- Relevante Informationen für die Entscheidungsfindung (Bell et al. 1988)

Ein Entscheidungsprozess kann mit rechnerbasierten Assistenzsystemen unterstützt werden (Buchholz und Clausen 2009). Diese zeichnen sich nach Buchholz und Clausen (2009) dadurch aus, dass Fakten dargestellt werden und die Lösung von Problemen sowie das Treffen von Entscheidungen erleichtert werden. Dabei kann zwischen zwei Arten von Unterstützungssystemen unterschieden werden. Einige Entscheidungsunterstützungssysteme schlagen dem Anwender Handlungsalternativen vor (Buchholz und Clausen 2009). Andere generieren keine Alternativen und präsentieren dem Anwender nur einen Handlungsvorschlag (Buchholz und Clausen 2009). Nach Buchholz und Clausen (2009) kann der Prozess einer Entscheidung in die Teilprozesse Entscheidungsvorbereitung, Entscheiden und Entscheidungsausführung gegliedert werden. Dies heißt, dass ein Assistenzsystem zur Entscheidungsunterstützung eine Lösungsmenge identifizieren, Alternativen wählen und bewerten sowie autonom agieren muss (Buchholz und Clausen 2009). Außerdem ist es wichtig, dass das Entscheidungsunterstützungssystem auf einer reproduzierbaren und systematischen Datenerhebung beruht (Kik 2022). Des Weiteren ist darauf zu achten, dass das Entscheidungsunterstützungssystem redundanzfrei und zerlegbar ist (Spengler et al. 2017). Wie bereits in der Einleitung erwähnt, ist eine Methode zur Entscheidungsunterstützung eine geeignete Möglichkeit die Auswahl einer Behandlungsmöglichkeit von Ausreißern zu unterstützen.

Das ist unter anderem darin begründet, dass die Methode zur Entscheidungsunterstützung verschiedene Handlungsmöglichkeiten übersichtlich darstellt.

2.2 Wissen

Daten bilden die Grundlage für Informationen und dem daraus folgenden Wissen. Der Begriff der Daten wird durch den ISO Standard ISO/IEC 2382-1 (2015) definiert. Demnach sind Daten eine reinterpretierbare Darstellung von Informationen in einer formalisierten Weise, die für die Kommunikation, Interpretation oder Verarbeitung geeignet ist.

In der Literatur wird zur Veranschaulichung bezüglich der Entstehung von Wissen durch Daten oft auf die Wissenspyramide verwiesen. Diese wird in der nachfolgenden Abbildung 1 abgewandelt beschrieben.

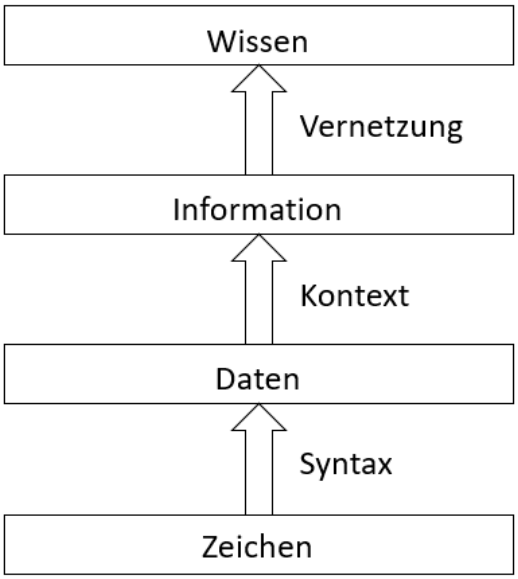

Abbildung 1 Wissensentstehung in Anlehnung an Bodendorf (2006), S. 1

In diesem Modell werden aufeinander aufbauende Ebenen verwendet, um die Zusammenhänge zwischen den Begriffen zu veranschaulichen. Wie der Abbildung 1 zu entnehmen ist, sind die Grundlage aller höheren Ebenen die Zeichen. Diese bilden mit Hilfe eines Zeichenvorrats und der Syntax nach definierten Regeln die Daten (Bodendorf 2006). Daten können in zwei Gruppen unterteilt werden. Zum einen gibt es Rohdaten. Diese Daten sind in ihrer einfachsten Form und noch nicht verarbeitet oder geprüft (Luber, Stefan und Litzel 2020), sie sind „ungeordnet, fehlerhaft, unvollständig, teilweise redundant oder unwichtig.“ (Runkler 2000, S. 2) Deshalb müssen Rohdaten in den meisten Fällen aufwendig angepasst werden, um im Anschluss Algorithmen auf die Daten anwenden zu können (Kureljusic und Karger 2022). Um Rohdaten handelt es sich zum Beispiel, wenn Temperaturmessungen vorgenommen werden oder Daten über den Verkehr erhoben werden. Dabei ist es unerheblich, ob die Aufzeichnung der Daten analog oder digital erfolgt.

Im Gegensatz dazu stehen die Sekundärdaten. Diese sind bereits mit Hilfe verschiedener Schritte und Methoden verarbeitet worden. Dazu zählt unter anderem die Erkennung und Behandlung von Fehlern sowie die Standardisierung der Daten.

Des Weiteren wird bei Daten zwischen den unstrukturierten und den strukturierten unterschieden. Bei unstrukturierten Daten ist oft nicht zu erkennen, ob Daten fehlen. Bei strukturierten Daten ist dies stets zu erkennen (Joenssen und Müllerleile 2014).

Daten, denen eine Bedeutung zugeordnet und die in einen Kontext gebracht wurden, werden zu Informationen (Bodendorf 2006). „Informationen ändern die Wahrnehmung des Empfängers in Bezug auf einen Sachverhalt und wirken sich auf die Beurteilung des Kontexts aus.“ (Bodendorf 2006, S. 1) In der heutigen Informatik werden Daten als Informationen bezeichnet, „die in eine binäre digitale Form umgewandelt wurden“ (Vaughan, Jack 2020). Informationen haben einige positive Eigenschaften. Sie können mit Algorithmen und Datenstrukturen übermittelt, klassifiziert und transformiert werden. Informationen unterliegen keinem physikalischen Alterungsprozess und sie benötigen keinen fixierten Träger. Einige negative Eigenschaften sind, dass zum Beispiel die Herkunft von einem einzelnen Teil der Informationen nicht nachweisbar ist. Das führt dazu, dass eine Manipulation nicht ausgeschlossen werden kann. Außerdem ist eine Information beliebig oft kopierbar und es kann kein Original identifiziert werden (Meier und Kaufmann 2016).

Durch die Vernetzung von verschiedenen Informationen entsteht das Wissen. Dazu sind Kenntnisse über den Zusammenhang der Informationen erforderlich. “Informationen sind sozusagen der Rohstoff, aus dem Wissen generiert wird, und die Form, in der Wissen kommuniziert und gespeichert wird.“ (North 2011, S. 37) Daten- und Informationsanalysen helfen dabei das Wissen zu vergrößern (Fasel und Meier 2016). Im Kontext von KDD und Data Mining werden unter dem Begriff Wissen interessante Muster verstanden, welche allgemein gültig sind, nicht trivial, neu, nützlich sowie verständlich (Runkler 2020).

2.3 Phasen der Wissensentdeckung

Die Wissensentdeckung in Datenbanken beschreibt „einen nicht trivialen Prozess, der Muster mit spezifischen Eigenschaften identifiziert.“ (Bensberg 2001, S. 65) Dabei muss zwischen den Begriffen des KDD Prozesses und des Data Mining unterschieden werden. Nach Fayyad et al. (1996) bezeichnet Data Mining die Anwendung von bestimmten Algorithmen mit deren Hilfe Muster aus Daten extrahiert werden können. Der KDD Prozess mit verschiedenen Phasen, zu denen auch das Data Mining zählt, soll sicher stellen, dass die gewonnenen Erkenntnisse brauchbar sind (Fayyad et al. 1996). Data Mining stellt demnach einen Bestandteil des KDD Prozesses dar. Eine Haupteigenschaft des KDD Prozesses ist es, ein iterativer Vorgang zu sein (Fayyad et al. 1996). Dies bezieht sich sowohl auf den gesamten Vorgang wie auch auf die enthaltenen Teilschritte (Sharafi 2012). Des Weiteren soll der Prozess durch umfassende Analysen neues Wissen aufzeigen.

Es gibt verschiedene Ansätze bezüglich der Phasen in der Wissensentdeckung in Datenbanken. Ein KDD Modell ist CRISP-DM (Cross Industry Standard Process for Data Mining). Dieses besteht aus sechs verschiedenen Phasen und wird von Wirth und Hipp (2000) wie nachfolgend beschrieben. Diese Phasen sind im Einzelnen das Domänenverständnis, das Datenverständnis, die Datenvorverarbeitung, die Modellierung, die Evaluierung und der Einsatz. Das CRISP-DM Modell eignet sich für Planung, Dokumentation und Kommunikation (Wirth und Hipp 2000). Im Folgenden werden die einzelnen Phasen erläutert.

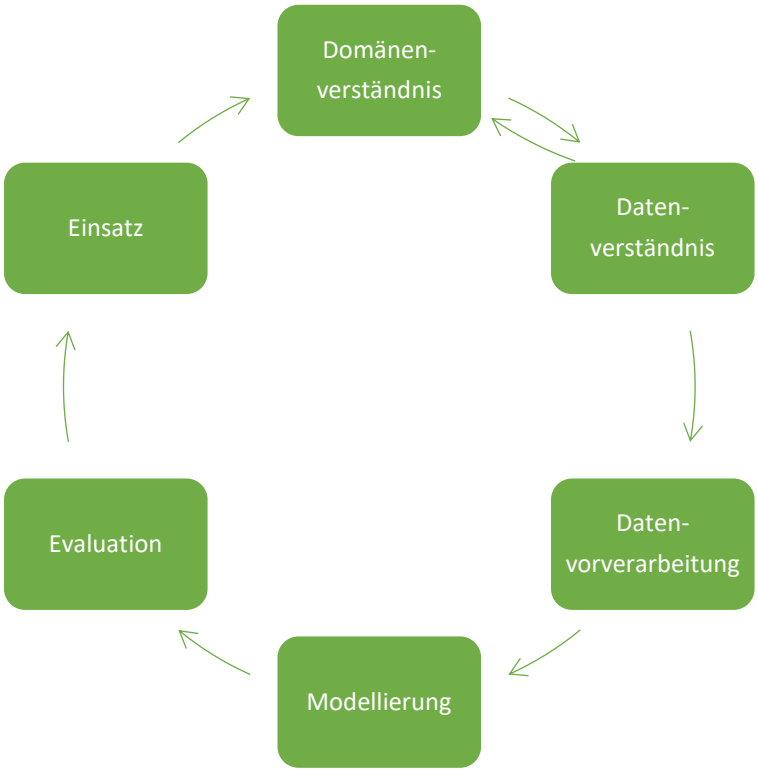

Abbildung 2 Phasen des CRISP-DM Prozesses in Anlehnung an Wirth und Hipp (2000), S. 5

In der ersten Phase des Domänenverständnisses steht das Verstehen der Ziele und Anforderungen des Projektes im Vordergrund. Außerdem soll in dieser Phase das Verstandene als ein Data Mining Problem formuliert werden. Des Weiteren wird eine erste Projektplanung vorgenommen (Wirth und Hipp 2000).

An die Phase des Domänenverständnisses schließt sich das Datenverständnis an. Hier werden die ersten Daten erhoben. Der Anwender macht sich in dieser Phase außerdem mit den Daten vertraut. Datenqualitätsprobleme werden erkannt und interessante Teilmengen, die eine Hypothesenbildung unterstützen, werden identifiziert. Zwischen den ersten beiden Phasen besteht eine enge Verknüpfung, da für die grobe Projektplanung und die Problemformulierung ein Verständnis der Daten erforderlich ist. Es ist möglich, von der Phase des Datenverständnis wieder zurück in die Phase des Domänenverständnis zu gelangen. Dies kann zum Beispiel passieren, wenn die Problemformulierung auf Grund des Datenverständnisses angepasst werden muss.

Nun folgt die Phase der Datenvorverarbeitung. Sie umfasst alle Vorgehensweisen die benötigt werden, um den finalen Datensatz zu erstellen (Wirth und Hipp 2000). Die vorliegenden Daten können aus strukturierten Daten, Mengen, Sequenzen, Texten, semi-strukturierten Texten, Zeitreihen, Graphen, Geo-Daten, Bildern, Audiodaten oder Videodaten bestehen (Müller und Lenz 2013). In diesem Schritt müssen unter anderem Fehler, Ausreißer und Rauschen entdeckt und entfernt oder reduziert werden (Runkler 2020). Des Weiteren sollten die Daten vektorisiert und normiert werden, sowie Merkmale extrahiert und fehlende Werte durch Imputationsverfahren plausibel geschätzt werden (Kureljusic und Karger 2022). Die Aufgaben in der Datenvorverarbeitung werden mehrfach und in keiner festgelegten Reihenfolge durchgeführt (Wirth und Hipp 2000). Wie bereits in der Einleitung beschrieben ist der Schritt der Datenvorverarbeitung für den gesamten Prozess der Wissensentdeckung wichtig, da er dazu beiträgt, unentdeckte und nützliche Informationen aus den Datenbeständen extrahieren zu können. Außerdem kann die Datenvorverarbeitung Einfluss auf die Qualität der späteren Data Mining Ergebnisse haben.

In der vierten Phase der Modellierung werden verschiedene Techniken sowohl ausgewählt als auch angewandt. Hier gibt es verschiedene Techniken, um ein Data Mining Problem behandeln zu können (Wirth und Hipp 2000). Das Ziel dieses Schrittes ist es, Wissen aus den vorhandenen Daten zu extrahieren (Runkler 2020). Durch Data Mining „kann eine hohe Qualität an Analyseergebnissen sichergestellt werden.“ (Joenssen und Müllerleile 2014, S. 459) Es bestehen einige Annahmen im Data Mining, die während des Prozesses getestet und kritisch hinterfragt werden sollten. Nach Müller und Lenz (2013) zählt dazu, dass Muster in der Vergangenheit auch noch in Zukunft gültig sein sollen. Des Weiteren sollen genügend Daten vorhanden sein sowie vorhandene Daten ausgewertet werden dürfen. Außerdem sollen diese Daten das enthalten, was prognostiziert werden soll (Müller und Lenz 2013). Zu den allgemeinen Aufgaben gehören die Segmentierung, die Abweichungsanalyse, die Klassifikation, die Prognose, die Assoziationsanalyse und die Sequenzanalyse (Müller und Lenz 2013). Eine Einteilung der Methoden zur Analyse kann nach dem Einsatzzweck vorgenommen werden.

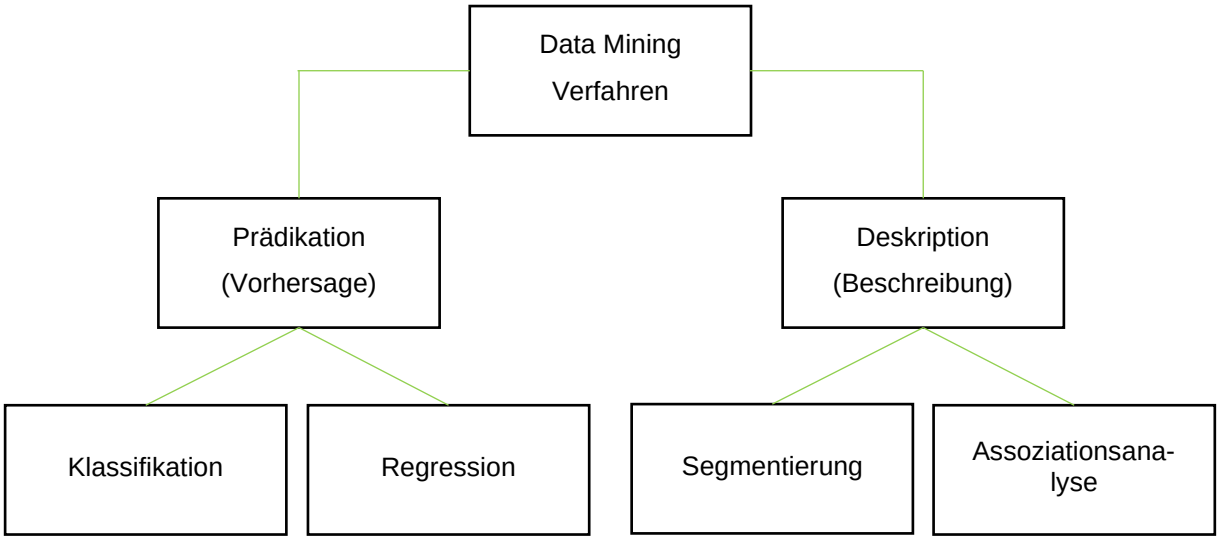

Abbildung 3 Übersicht über Analyseverfahren im Data Miningin Anlehnung an D'Onofrio und Meier (2021), S. 27

Wie in Abbildung 3 zu sehen ist, unterscheiden unter Anderem D’Onofrio und Meier (2021) zwischen Verfahren zur Prädikation und Deskription. Die Prädikationsverfahren können als überwacht beschrieben werden während die Deskriptionsverfahren als unüberwacht gelten. Die überwachten Verfahren verwenden einen historischen Datenbestand mit bekannten Ergebnissen. Sie streben es an, den Einfluss der Ausprägungen von unabhängigen Variablen auf die Ausprägung abhängiger Variablen abzuschätzen. Das angewandte Verfahren „‘lernt‘ somit, wie sich aus einer vorgegebenen Kombination der Ausprägungen von Eingangsvariablen Ausgabe- oder Ergebniswerte ermitteln lassen.“ (D'Onofrio und Meier 2021, S. 28) Neben Entscheidungsbäumen und künstlichen neuronalen Netzen gehören zu den überwachten Verfahren auch die Klassifikation und die Regression. Zu den unüberwachten Verfahren gehören unter anderem die Clusteranalysen und die Assoziationsanalysen. Die unüberwachten Verfahren suchen nach Mustern in den Ausprägungen von Variablen, ohne diese Variablen zu verwenden. Sie können keine Prognose über die zu beobachtenden Ausprägungen erstellen (D'Onofrio und Meier 2021). Analysen durch Data Mining sind besonders für Problemstellungen geeignet, bei welchen eine wissensbasierte und komplexe Entscheidung verlangt wird oder eine hinreichende Masse and Daten vorliegt. Probleme, die zum aktuellen Zeitpunkt mit suboptimalen Methoden bewältigt werden sind ebenfalls gut für das Data Mining geeignet (Müller und Lenz 2013). Die Phasen der Datenvorverarbeitung und der Modellierung haben

einen engen Bezug zueinander. Zum Beispiel kann es vorkommen, dass während der Modellierung Datenprobleme erkannt werden (Wirth und Hipp 2000). Diese Phase wird im folgenden Verlauf dieser Arbeit als das Data Mining bezeichnet. Wie bereits in der Einleitung erwähnt, trennen einige Autoren die Begriffe des KDD und Data Mining nicht. In dieser Arbeit wird das Data Mining als Teil des KDD Prozesses betrachtet.

In der Phase der Evaluation ist es wichtig, die vorher verwendeten Modelle zu bewerten und die durchgeführten Schritte zu überprüfen. Dies soll sicherstellen, dass die Ziele erreicht werden können. Hier ist besonders darauf zu achten, dass kein Kernproblem unbeachtet geblieben ist und kein Aspekt nicht ausreichend berücksichtigt wurde (Wirth und Hipp 2000). Außerdem sollen alle vorangegangenen Schritte mit in die Interpretation einfließen, um einen sachlogischen Begründungszusammenhang herstellen zu können (Bensberg 2001). Die Erkenntnisse sollen in diesem Schritt außerdem mit vorher verfügbarem Wissen abgeglichen und bewertet, sowie auf die Einsetzbarkeit geprüft werden (Müller und Lenz 2013). Zum Ende dieser Phase sollte festgestellt werden können, dass die Ergebnisse des Data Mining nützlich sind. Das CRISP-DM Modell schließt mit der Phase des Einsatzes ab. Hier wird das erlangte Wissen so präsentiert, dass es genutzt werden kann. Die Anforderungen an die Präsentation sind projektabhängig. Oft werden die Implementierungsschritte nicht von einem Datenanalysten, sondern vom Anwender durchgeführt. Es ist darauf zu achten dass im Vorfeld klar ist, welche Schritte durchgeführt werden müssen. Um das erste Teilziel vollständig zu erreichen, wird neben der bereits beschriebenen Entstehung von Wissen und einer Übersicht über den KDD Prozess im Folgenden genauer auf die Datenvorverarbeitung und Ausreißer eingegangen, da diese unter anderem die Grundlage für den weiteren Verlauf dieser Arbeit darstellen.

2.4 Datenvorverarbeitung

Wie in Abschnitt 2.2 beschrieben, ist die Datenvorverarbeitung ein wichtiger Bestandteil von KDD Prozessen. Rohdaten sind zunächst unvollkommen und enthalten Inkonsistenzen sowie Redundanzen (Luengo et al. 2020). Große Mengen an Daten müssen mit anspruchsvollen Mechanismen analysiert werden, weshalb sie vor der Anwendung des Data Mining vorverarbeitet und angepasst werden sollten (Luengo et al. 2020). Die Qualität der Daten nach der Datenvorverarbeitung spielt eine große Rolle für das anschließende Data Mining (Alasadi und Bhaya 2017). Data Mining Modelle können bei ausreichender Datenqualität mit geringerem Zeitaufwand und einer größeren Genauigkeit angewandt werden (Luengo et al. 2020). Ziel der Datenvorverarbeitung ist es einen Datensatz zu schaffen, der als zuverlässige und geeignete Ausgangslage für die Anwendung des Data Mining dient (Luengo et al. 2020). Nützliche Informationen sollen so für eine effiziente Modellierung genutzt werden können (Mishra et al. 2020). Die spezifischen Ziele der Datenvorverarbeitung hängen immer von der Art der zu behandelnden Fehler ab (Mishra et al. 2020). Außerdem kann die Datenvorverarbeitung dazu beitragen, bestehende Data Mining Modelle zu verbessern (Luengo et al. 2020). Zu den Bestandteilen der Datenvorverarbeitung zählen unter anderem die Datentransformation, Integration, die Bereinigung der Daten sowie die Normalisierung (Luengo et al. 2020). Die Datensätze mit welchen gearbeitet wird, enthalten aus verschiedenen Gründen Fehler. Sie beinhalten fehlende Werte, verrauschte oder unvollständige Daten sowie Ausreißer (Alasadi und Bhaya 2017). Oft werden für die Behandlung eines Fehlers mehrere Verfahren angewandt, um mit dem Fehler effektiv umgehen zu können (Mishra et al. 2020).

Im Folgenden wird kurz auf fehlende Daten eingegangen. Im Anschluss daran werden die Begriffe des Rauschens und der Ausreißer erläutert, wobei eine genauere Beschreibung von Ausreißern in dem Abschnitt 2.5 erfolgt.

Allgemein können fehlende Daten mehrere Ursachen haben (Mishra et al. 2020). Dazu zählen Werte, die außerhalb des Messbereiches liegen oder verschiedene Sensoren mit einer unterschiedlichen Abtastrate (Mishra et al. 2020). Mit Hilfe von Datenimputationsverfahren können diese fehlenden Werte geschätzt werden (Mishra et al. 2020).

Nach Runkler (2020) können Fehler in Kategorien unterteilt werden. Mess- und Übertragungsfehler lassen sich den zufälligen Fehlern zuordnen. Sie können als Ausreißer vorliegen oder als additives Rauschen abgebildet werden. Als Rauschen werden Fehler bezeichnet, welche zufällig aufgezeichnet werden und die den Wert einer Variable verändern (Aggarwal 2015). Dabei lässt sich das Rauschen in zwei weitere Kategorien unterteilen. Das Klassenrauschen bezeichnet Daten, deren Beschriftung fehlerhaft ist (García et al. 2015). Attributrauschen liegt vor, wenn ein oder mehrere Attributwerte verfälscht wurden (García et al. 2015). Eine Ursache für Rauschen kann zum Beispiel die Empfindlichkeit eines Messinstrumentes sein (Mishra et al. 2020).

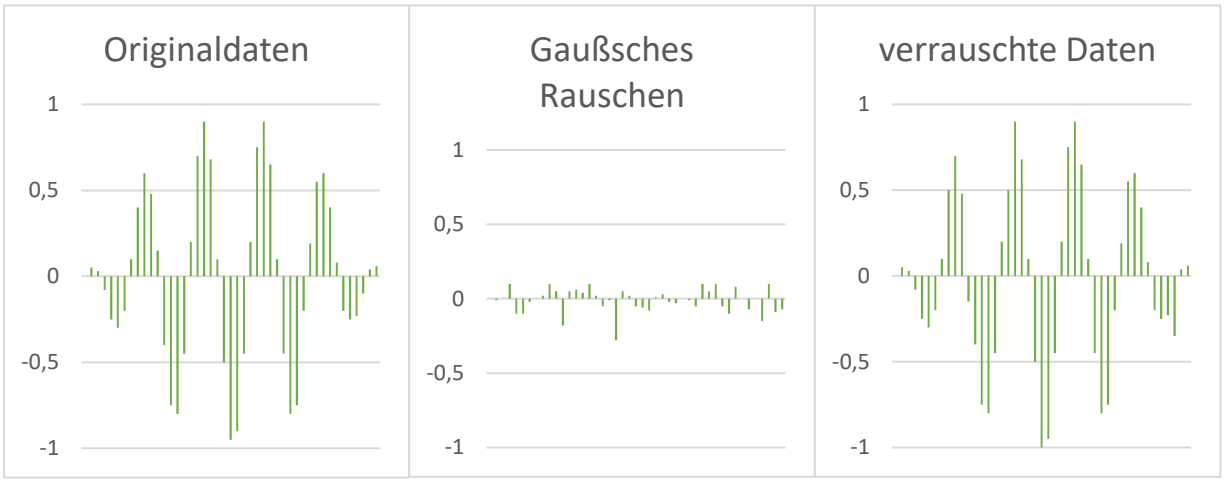

Abbildung 4 Originaldaten, Gaußsches Rauschen und verrauschte Daten in Anlehnung an Runkler (2020), S. 24

In Abbildung 4 sind Originaldaten, Gaußsches Rauschen und verrauschte Daten zu sehen. Dabei ergeben die Originaldaten zusammen mit dem Gaußschen Rauschen die verrauschten Daten. Es ist zu erkennen, dass der verrauschte Datensatz qualitativ nur einen geringfügigen Unterschied zu dem Originaldatensatz aufzeigt. Daher hat moderates Rauschen einen geringen Einfluss auf die Datenanalyse (Runkler 2020). Wird der Originaldatensatz durch Rauschen stärker verändert, muss dieses nach Möglichkeit verringert bzw. entfernt werden (Fayyad et al. 1996). Dies kann manuell erfolgen oder zum Beispiel durch die Verwendung von Filteralgorithmen (Mishra et al. 2020). Außerdem können Glättungsfunktionen oder eine Regression durchgeführt werden (Alasadi und Bhaya 2017).

Die zweite Kategorie, in die Fehler eingeordnet werden können, sind die systematischen Fehler. Diese können zum Beispiel durch fehlerhafte Formeln bei der Berechnung verursacht werden. Wenn die Systematik des Fehlers bekannt ist, kann dieser korrigiert werden (Runkler 2020). Nach Runkler (2020) gehören hierzu unter anderem Messfehler, die durch eine falsche Skalierung oder Drifteffekte ausgelöst werden. Außerdem können laut Runkler durch Verarbeitungs- oder Erfassungsfehler Ausreißer hervorgerufen werden. Diese Fehler treten bei der manuellen Datenerfassung häufiger auf als bei der automatischen, zum Beispiel weil einzelne Ziffern vertauscht werden (Runkler 2020).

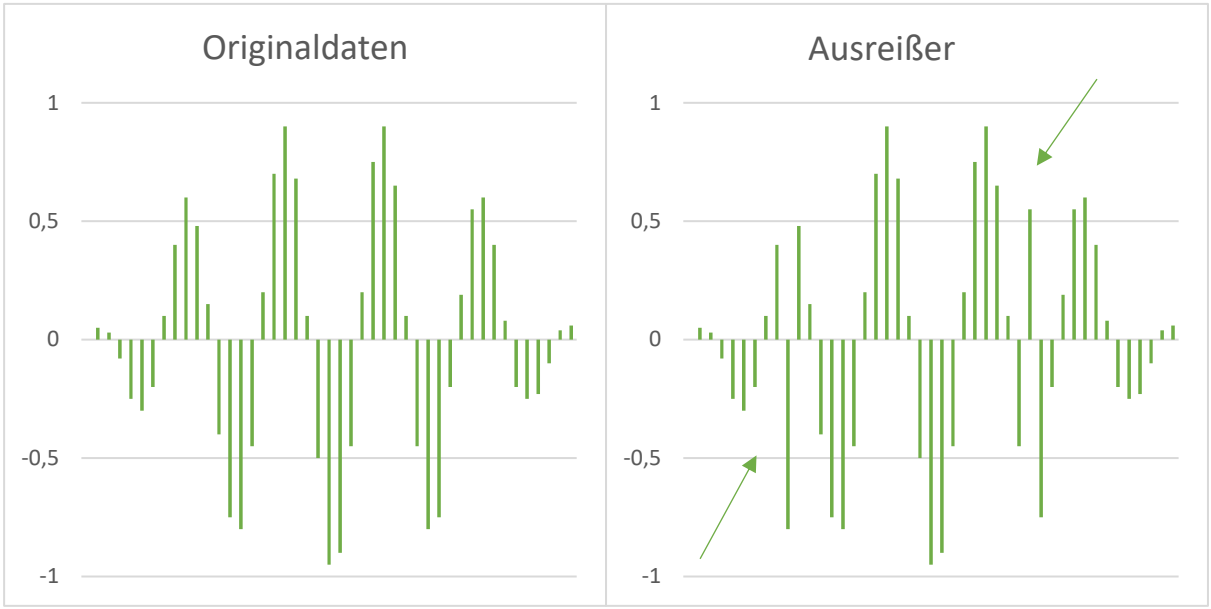

Abbildung 5 Originaldaten und Ausreißerin Anlehnung an Runkler (2020), S. 24

Durch Abbildung 5 wird deutlich, dass die Ausreißeridentifizierung herausfordernd ist. Datensätze müssen genau untersucht werden, um die Ausreißer klar identifizieren und behandeln zu können (Runkler 2000). Jeder Schritt in der Datenvorverarbeitung kann Risiken, zum Beispiel durch falsche Anwendung, für den ganzen KDD-Prozess bedeuten. Daher sollte der Datenvorverarbeitung eine hohe Aufmerksamkeit beigemessen werden, damit der gesamte Prozess erfolgreich durchgeführt werden kann (Runkler 2000). Die ausgewählten Datenvorverarbeitungsverfahren müssen immer an die Beschaffenheit der Daten angepasst werden und beeinflussen den Erfolg des kompletten KDD Prozesses (Mishra et al. 2020). Automated Machine Learning (AutoML)-Tools können bei einigen Schritten unterstützen, wie zum Beispiel in der Vektorisierung von kategorialen Variablen (Kureljusic und Karger 2022). Das ersetzt jedoch keine vollständige Datenvorverarbeitung durch einen Analysten, da besonders in der Merkmalsextraktion ein umfangreiches Verständnis der Datenmengen und deren Zusammenhängen notwendig ist (Kureljusic und Karger 2022). In der Literatur werden Ausreißer und Rauschen teilweise nicht getrennt betrachtet und die Begriffe synonym verwendet. Im Verlauf dieser Arbeit wird zu erkennen sein, dass sich einige Verfahren für die Behandlung von Ausreißern und Rauschen eignen und andere nur für die Behandlung von Ausreißern. Aus diesem und den oben genannten Gründen ist es wichtig, die Begriffe Ausreißer und Rauschen getrennt voneinander zu betrachten. Dadurch soll erreicht werden, dass die Behandlungsverfahren bezüglich Ausreißern in der Datenvorverarbeitung genau auf die vorliegenden Daten angepasst sind.

2.5 Ausreißer

Wie in Abschnitt 2.3 erwähnt, kann die Ausreißeridentifizierung herausfordernd sein. Wie bereits in der Einleitung beschrieben, kann man im Allgemeinen unter dem Begriff des Ausreißers alle Datenpunkte eines Datenbestandes zusammenfassen, welche im Gegensatz zu den anderen Datenpunkten deutlich abweichen bzw. auffällig sind (S. Baumann et al. 2018). Hawkins (1980) erwähnt zusätzlich, dass bei diesen Abweichungen der Verdacht entstehen könnte, dass die Beobachtungen durch einen anderen Mechanismus entstanden sei. Nach Aggarwal (2017) können diese abweichenden Werte in der Literatur auch als Abnormalitäten, Abweichungen und widersprüchliche Beobachtungen bezeichnet werden. In den meisten Anwendungen werden die Daten durch einen oder mehrere Generierungsprozesse erzeugt, die Aktivitäten im System widerspiegeln können (Aggarwal 2017). Wenn sich der Erzeugungsprozess ungewöhnlich verhält, kann das dazu führen, dass Ausreißer erzeugt werden (Aggarwal 2017). Daher enthalten Ausreißer oft nützliche Informationen über ungewöhnliche Merkmale in den Systemen (Aggarwal 2017). Daraus können sich unter anderem Rückschlüsse auf den Prozess der Datenerzeugung ziehen lassen (Aggarwal 2017). Aggarwal (2017) nennt einige Beispiele, in denen die Ausreißer nützliche Informationen teilen. Zum Beispiel kann ein Kreditkartenbetrug dadurch aufgedeckt werden, dass plötzlich hohe Transaktionen von anderen Orten aus vorgenommen werden die nicht zu dem sonstigen Bezahlverhalten passen (Aggarwal 2017). Einen bedeutenden Aspekt der Ausreißererkennung und -behandlung stellt die Art der Daten dar (Chandola et al. 2009). Dateninstanzen können aus einem oder mehreren Attributen bestehen (Chandola et al. 2009). Die Attribute der Dateninstanzen können aus verschiedenen Typen bestehen, wie zum Beispiel binär, kategorial oder kontinuierlich (Chandola et al. 2009). Falls eine Dateninstanz aus mehreren Attributen besteht, können mehrere oder nur ein Datentyp vorliegen (Chandola et al. 2009). Die Art der Attribute bestimmt die Anwendbarkeit von Techniken zur Erkennung und Behandlung von Ausreißern, zum Beispiel müssen verschiedene Modelle für kontinuierliche und kategoriale Daten verwendet werden (Chandola et al. 2009). Es kann zwischen lokalen und globalen Ausreißern unterschieden werden. Lokale Ausreißer liegen innerhalb der Verteilung der übrigen Werte (Hawkins 1980). Daher sind sie besonders schwer zu erkennen (Runkler 2020). Globale Ausreißer hingegen liegen außerhalb der Verteilung der restlichen Werte und weichen damit deutlich ab (Hawkins 1980). Lokale und globale Ausreißer lassen sich, wie bereits in der Einleitung erwähnt, nach Chandola et al. (2009) in drei Hauptkategorien unterteilen. Diese werden als punktuelle, kontextuale und kollektive Ausreißer bezeichnet (Chandola et al. 2009).

Ein punktueller Ausreißer liegt dann vor, wenn eine einzelne Dateninstanz im Vergleich zu den restlichen Instanzen abweicht (Zhang et al. 2020). Dies kann als die einfachste Art eines Ausreißers bezeichnet werden und steht in vielen Forschungsarbeiten im Fokus (Divya und Babu 2016). Sie werden allgemein im Zusammenhang mit mehrdimensionalen Datentypen untersucht (Gupta et al. 2014). Punktausreißer können sowohl mit unüberwachten Verfahren als auch mit überwachten Verfahren erkannt und behandelt werden (Aggarwal 2017; Rajeswari et al. 2018).

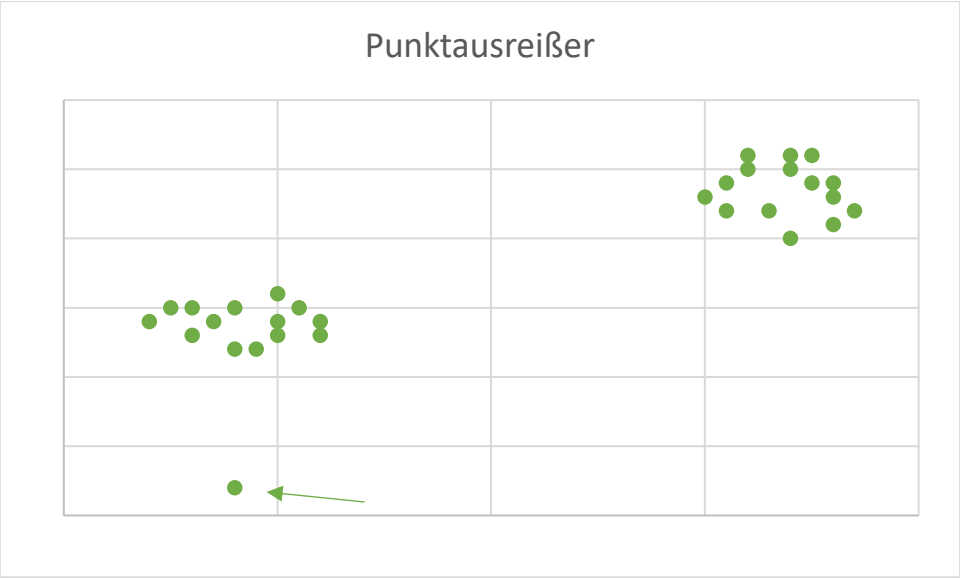

Abbildung 6 Darstellung eines Punktausreißer in Anlehnung an Ranga Suri et al. (2019), S. 4

In Abbildung 6 wird ein Punktausreißer dargestellt. Er weicht deutlich von den zwei Clustergruppen ab (Ranga Suri et al. 2019). Neben den Punktausreißern gibt es auch die kontextualen Ausreißer. Davon wird gesprochen, wenn Dateninstanzen in einem bestimmten Kontext als anormal bezeichnet werden können, jedoch nicht in einer anderen Art (Divya und Babu 2016). Der Begriff des Kontextes muss in diesem Fall in der Problemformulierung spezifiziert werden (Chandola et al. 2009). Jede Instanz kann mit den folgenden zwei Attributen definiert werden. Kontextuelle Attribute werden verwendet, um den Zusammenhang zwischen den Dateninstanzen zu bestimmen (Chandola et al. 2009). Im räumlichen Zusammenhang kann dies beispielsweise eine Längengradangabe sein (Chandola et al. 2009). Verhaltensattribute definieren nicht-kontextuelle Merkmale (Chandola et al. 2009). Im räumlichen Zusammenhang könnte dies zum Beispiel der Niederschlag an einem bestimmten Ort sein (Chandola et al. 2009). Das anormale Verhalten zeigt sich also anhand der Verhaltensattribute innerhalb eines bestimmten Kontextes (Chandola et al. 2009). Eine Dateninstanz kann demnach in dem einen Kontext eine Anomalie darstellen und in einem anderen Kontext als normal bezeichnet werden (Chandola et al. 2009).

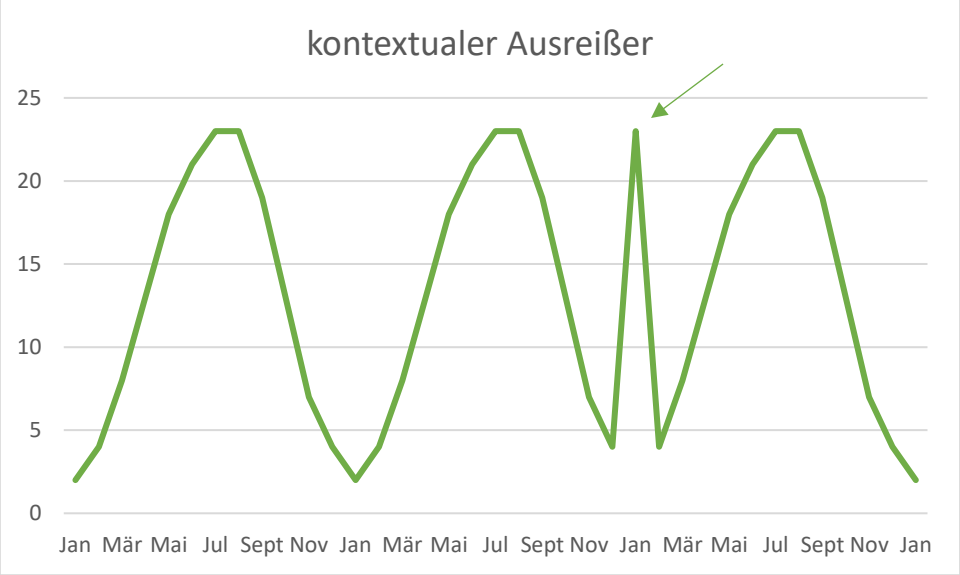

Abbildung 7 Darstellung eines kontextualen Ausreißers in Anlehnung an Chandola et al. (2009), S. 8

In Abbildung 7 ist exemplarisch der Temperaturverlauf in Deutschland zu sehen. Während Temperaturen im Sommer von 23 °C normal sind, ist diese Temperatur im Winter eine Anomalität. Daher ist dies als kontextueller Ausreißer anzusehen, da es nicht der gewöhnlichen Temperatur im Kontext der Jahreszeit entspricht. Die Auswahl eines Behandlungsverfahrens von kontextualen Ausreißern ist davon abhängig, wie bedeutsam diese Ausreißer in Bezug auf den Zielanwendungsbereich ist (Chandola et al. 2009). Sie werden im Zusammenhang mit abhängigkeitsorientierten Datentypen untersucht (Gupta et al. 2014). Nach Chandola et al. (2009) und Gupta et al. (2014) liegen kontextuale Ausreißer häufig in Zeitreihen und räumlichen Diagrammen vor. Außerdem ist die Möglichkeit einer Definition des Kontextes ausschlaggebend, ob ein Behandlungsverfahren angewendet werden kann (Chandola et al. 2009). Im Allgemeinen können kontextuale Ausreißer mit unüberwachten Verfahren, wie zum Beispiel der Clusteranalyse, behandelt werden (Rajeswari et al. 2018).

Die dritte Kategorie, in die Ausreißer unterteilt werden können, sind die kollektiven Ausreißer (Divya und Babu 2016). Hier handelt es sich um eine Ansammlung von Dateninstanzen, die für sich genommen nicht zwingend Ausreißer sein müssen, allerdings ist ihr gemeinsames Auftreten in einer Sammlung anormal (Chandola et al. 2009). Kollektive Ausreißer treten nur in Datensätzen auf, in denen die Dateninstanzen miteinander verbunden sind (Divya und Babu 2016). Die Datensätze können unter anderem aus Sequenzdaten, Graphdaten und räumlichen Daten bestehen (Chandola et al. 2009). Im Allgemeinen sind überwachte Verfahren geeignet, um kollektive Ausreißer zu behandeln (Rajeswari et al. 2018). Dazu gehören neben weiteren die Klassifikationsverfahren (D'Onofrio und Meier 2021).

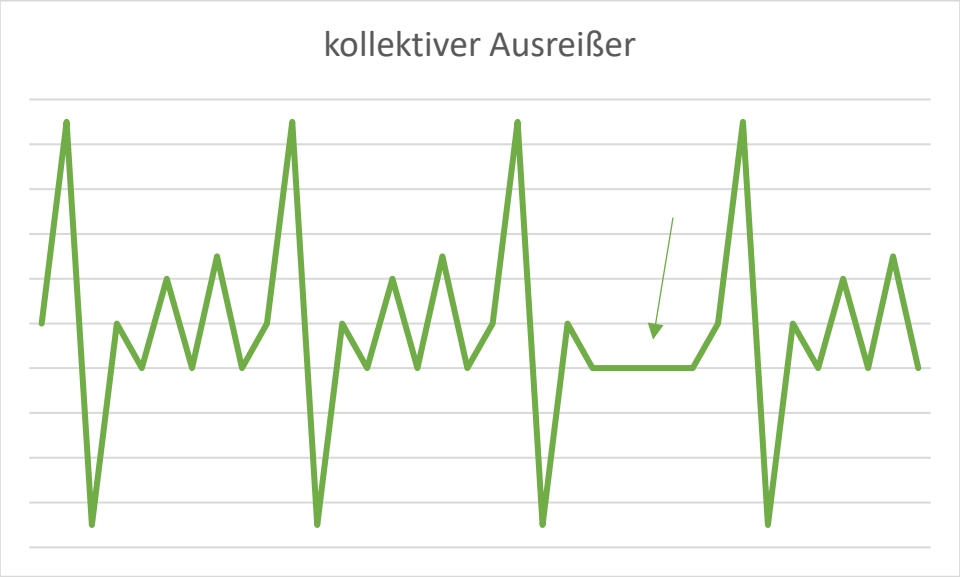

Abbildung 8 Darstellung eines kollektiven Ausreißers in Anlehnung an Chandola et al. (2009), S. 9

In Abbildung 8 ist ein kollektiver Ausreißer zu sehen. Es ist zu erkennen, dass die einzelnen Werte für sich gesehen nicht von den Übrigen abweichen. Das gemeinsame Auftreten jedoch stellt eine Anomalie dar. Kollektive Ausreißer können ebenfalls in Zeitreihen sowie in Bilddatensätzen vorliegen (Aggarwal 2017). Generell können kollektive Ausreißer mit überwachten Verfahren behandelt werden (Rajeswari et al. 2018).

Zusammenfassend lässt sich sagen, dass punktuelle Ausreißer in jedem Datensatz vorkommen können, kontextuelle Ausreißer von der Verfügbarkeit von Kontextattributen abhängen und kollektive Ausreißer nur in Datensätzen vorkommen, in denen Beobachtungen miteinander verbunden sind. Außerdem können punktuelle und kollektive Ausreißer auch kontextuelle Ausreißer sein, wenn sie in Bezug auf einen bestimmten Kontext analysiert werden. Nach

Aggarwal (2017) ist ein Ausreißer in zeitreihen kontextual, wenn er einzeln vorliegt. Die Unterscheidung zwischen Punkt-, kontextualem und kollektivem Ausreißer ist wichtig, da nicht alle Behandlungsverfahren von Ausreißern in der Datenvorverarbeitung jede der genannten Hauptkategorien erkennen und damit behandeln können.

Nachdem ein Ausreißer identifiziert wurde besteht der nächste Schritt innerhalb der Datenvorverarbeitung darin, sich mit seinem Vorhandensein zu befassen sowie unter Umständen seine Auswirkungen auf den Datensatz zu behandeln (Romão und Vasanelli 2021). Durch zusätzliche Messungen kann überprüft werden, ob der Ausreißer auf einen eliminierbaren Fehler zurückzuführen ist (Romão und Vasanelli 2021). Die Entscheidung, wie mit Ausreißern umgegangen wird, hat einen großen Einfluss auf inhaltliche Schlussfolgerungen (Aguinis et al. 2013). Die Entscheidungen zur Verfahrensauswahl sollten detailliert beschrieben werden um zu gewährleisten, dass genügend Transparenz und Nachvollziehbarkeit vorherrscht (Aguinis et al. 2013). Die meisten Behandlungsverfahren von Ausreißern können in eine der folgenden drei Möglichkeiten eingeordnet werden (Romão und Vasanelli 2021). Wurde ein Ausreißer als Fehler identifiziert, kann diese Beobachtung gegebenenfalls durch einen korrekten Wert ersetzt werden (Romão und Vasanelli 2021). Ghavami (2020) schlägt für diese erste Möglichkeit Regressionsverfahren vor. Besteht eine fundierte Begründung, dass ein Ausreißer ein Fehler ist und nicht durch einen korrekten Wert ersetzt werden kann, ist es möglich, diesen Ausreißer zu entfernen (Romão und Vasanelli 2021). Für diese Möglichkeit eignen sich nach Ester und Sander (2000) Clusteranalysen. Die dritte Möglichkeit besteht darin einen Ausreißer so zu berücksichtigen, dass seine Auswirkung auf die nachfolgende Analyse verringert wird (Romão und Vasanelli 2021). Für diese Möglichkeit eignet sich zum Beispiel das Behandlungsverfahren des Winsorisierens (Aguinis et al. 2013). Die hier genannten Behandlungsverfahren werden in Kapitel 3 genauer erläutert. Ein weiteres Verfahren zur Behandlung von Ausreißern ist nach Runkler (2020) die Klassifikation. Bei einigen der genannten Verfahren können sogenannte Metaalgorithmen verwendet werden, um die Robustheit der Ergebnisse zu verbessern (Aggarwal 2017). Sie kombinieren die Ergebnisse verschiedener Algorithmen und werden auch als Ensembles bezeichnet (Aggarwal 2017). Die Datenpunkte können auf Grundlage der Ergebnisse der verschiedenen Algorithmen bewertet werden, was als Ensemble-Score bezeichnet werden kann (Aggarwal 2017). Dies kann dabei unterstützen, Ausreißer zu behandeln (Aggarwal und Sathe 2017). In der Literatur gibt es eine Vielzahl an weiteren Behandlungsmöglichkeiten. Es besteht zum Beispiel die Möglichkeit, eine Fehlerliste zu führen (Runkler 2020). In diesem Fall bleiben Fehler, zu denen nach Runkler (2020) auch die Ausreißer zählen, unverändert. Allerdings werden die Indizes der Daten, die Fehler enthalten, in einer separaten Liste gespeichert (Runkler 2020). Außerdem ist es möglich, ungültige Daten, zu denen auch Ausreißer zählen können, durch einen speziellen Fehlerwert zu ersetzen (Runkler 2020). Nach Cleve und Lämmel (2020) bietet sich bei Ausreißern der gleiche Umgang wie mit fehlenden Werten an. Hier besteht eine Möglichkeit darin, das Attribut zu ignorieren (Cleve und Lämmel 2020). Dabei wird eine ganze Spalte aus der Datentabelle gelöscht, um das Attribut zu entfernen (Cleve und Lämmel 2020). Cleve und Lämmel (2020) weisen allerdings darauf hin, dass ein Attribut nur dann gelöscht werden darf, wenn eine genügend große Anzahl an anderen, vollständigen Attributen vorliegt. Nach Aguinis et al. (2013) eignet sich die Bayessche Statistik ebenfalls zur Ausreißerbehandlung. Ein Algorithmus, welcher auf dieser Statistik beruht, ist der Naive-Bayes-Algorithmus (Cleve und Lämmel 2020). Hier wird die grundlegende Annahme getroffen, dass alle Attribute unabhängig voneinander sind (Cleve und Lämmel 2020). Außerdem wird direkt aus Trainingsdaten eine Vorhersage berechnet (Cleve und Lämmel 2020). Dieses Verfahren lässt sich zu den Klassifikationsverfahren zuordnen. Einzelne Verfahren zur Behandlung von Ausreißern in der Datenvorverarbeitung werden in Kapitel 3 nach der strukturierten Literaturrecherche genauer vorgestellt.

Mit Hilfe der in diesem Kapitel behandelten Abschnitte ist das erste Teilziel der Arbeit erreicht. Es wurde ein allgemeiner Überblick über Methoden zur Entscheidungsunterstützung gegeben sowie die Entstehung von Wissen erläutert. Außerdem wurden Grundlegendes in Bezug auf

einen KDD Prozess erläutert, sowie die Begriffe des Data Mining und der Datenvorverarbeitung beschrieben. Im letzten Abschnitt dieses Kapitels wurden verschiedene Arten von Ausreißern definiert und Möglichkeiten zur Behandlung von Ausreißern vorgestellt. Auf dieser Grundlage können wichtige Schlüsselbegriffe abgeleitet werden, die für die strukturierte Literaturrecherche aus dem nachfolgenden Kapitel 3 verwendet werden.

3 Verfahren zur Behandlung von Ausreißern

In diesem Kapitel werden verschiedene Verfahren zu der Behandlung von Ausreißern in der Datenvorverarbeitung in Vorbereitung auf das Data Mining vorgestellt. Um Literatur zu ermitteln, welche die Verfahren in ihrer Vorgehensweise beschreiben, wird das Vorgehen einer strukturierten Literaturrecherche nach vom Brocke et al. (2009) angewandt. Dieses wird im Folgenden in der Theorie erläutert sowie die Anwendung auf diese Arbeit definiert. Im anschließenden Teil dieses Kapitels werden die ermittelten Verfahren vorgestellt und die Ansichten der Autoren bezüglich der Eignung zur Behandlung von Ausreißern gegenübergestellt.

Die Entdeckung von neuem Wissen kann durch die Kombination und Interpretation von bereits vorhandenem Wissen entstehen (Vom Brocke et al. 2009). Nach vom Brocke et al. (2009) spielen Literaturübersichten daher eine wichtige Rolle in der Wissenschaft. Die Qualität dieser wird maßgeblich von der Qualität der Literatursuche bestimmt (Vom Brocke et al. 2009). Behandlungsmöglichkeiten von Ausreißern in der Datenvorverarbeitung sollen in dieser Bachelorarbeit auf der Grundlage bereits bestehender Behandlungsmöglichkeiten vorgestellt werde. Um diese erkennen zu können wird eine systematische Literaturrecherche durchgeführt. Im Folgenden wird das Konzept und die Durchführung der systematischen Literaturrecherche nach vom Brocke et al. (2009) erläutert. Das Ziel des Verfassens von einer Literaturübersicht besteht darin, gesammeltes Wissen eines Bereiches zusammenzutragen und zu rekonstruieren. Die Grundlage hierfür stellt die strukturierte Literaturrecherche dar (Vom Brocke et al. 2009). Diese zielt darauf ab, relevante Quellen für ein Thema zu finden. Sie leistet damit einen großen Beitrag zu Relevanz und Strenge in der Forschung (Vom Brocke et al. 2009). Die Relevanz verbessert sich dadurch, dass die Recherche von bereits bekannten Quellen vermieden wird. Die Strenge leitet sich aus einer effektiven Nutzung der bereits bestehenden Wissensbasis ab (Vom Brocke et al. 2009). Der Prozess der strukturierten Literaturrecherche umfasst einerseits die Identifizierung von qualitativ hochwertigen Quellen und andererseits die Bewertung ihrer Anwendbarkeit (Vom Brocke et al. 2009).

Abbildung 9 Phasen einer strukturierten Literaturrecherche in Anlehnung an vom vom Brocke et al. (2009), S. 7

Wie in Abbildung 9 zu sehen ist, besteht die strukturierte Literaturrecherche nach vom Brocke et al. (2009) aus fünf Phasen. Diese sind die Definition des Rechercheumfangs, die Konzepterstellung, die Literaturrecherche, die Literaturanalyse und die Forschungsagenda.

In der ersten Phase „Definition des Rechercheumfangs“ wird der Umfang der Literaturrecherche festgelegt und die Forschungsfrage definiert. Um den Umfang klar definieren zu können, schlagen vom Brocke et al. (2009) vor auf die von Cooper (1988) vorgestellte und etablierte

Taxonomie für Literaturrecherchen zurückzugreifen. Diese Taxonomie besteht auch sechs Merkmalen die im Folgenden dargestellt und genannt werden.

Tabelle 1 Taxonomie einer Literaturrecherche in Anlehnung an Cooper (1988), S. 109

Merkmal	**Kategorie**
Fokus	Forschungsergebnisse, Forschungsmethoden, Theorien, Praktiken, Anwendungen
Ziele	Integration, Kritik, Identifikation zentraler Aspekte
Organisation	historisch, methodisch, konzeptionell
Perspektive	neutrale Darstellung, Vertretung eines Standpunktes
Zielgruppe	spezialisierte Wissenschaftler, allgemeine Wissenschaftler, öffentliches Publikum
Abdeckungsgrad der Quellen	vollständig, vollständig selektiv, repräsentativ, zentrale Abdeckung

Wie in Tabelle 1 zu sehen ist, ist das erste Merkmal der Fokus der Literaturrecherche. Er liegt auf dem, was der Verfasser präsentieren möchte (Cooper 1988). Die Ziele sind das Zusammenfassen, Kritisieren und Integrieren von Erkenntnissen und bilden das zweite Merkmal (Cooper 1988). Das dritte Merkmal ist die Organisation. Hier wird von Cooper (1988) vorgeschlagen, eine historische, methodische oder konzeptionelle Struktur zu wählen. Die Perspektive einer Quelle stellt das vierte Merkmal dar. Hier soll festgelegt werden, ob eine bestimmte Position vertreten wird oder eine neutrale Darstellung erfolgt (Cooper 1988). Die Zielgruppe stellt das fünfte Merkmal dar. Die Zielgruppe können beispielsweise spezialisierte Wissenschaftler oder ein öffentliches Publikum sein (Cooper 1988). Das letzte Merkmal stellt den Abdeckungsgrad der Quellen dar. Dieser kann beispielsweise vollständig oder repräsentativ sein (Cooper 1988). Die Anwendung dieser Taxonomie ist nach vom Brocke et al. (2009) notwendig und hat Auswirkungen auf den anschließenden Suchprozess. Mit Hilfe der genannten Merkmale soll die Arbeit außerdem klassifiziert werden (Vom Brocke et al. 2009).

In der zweiten Phase „Konzepterstellung“ soll die Analyse organisiert werden. Zum Beispiel sollen die Datenbanken definiert werden und Arbeitsdefinitionen der Schlüsselbegriffe gegeben werden. Es ist wichtig eine Übersicht über bereits vorhandenes Wissen zu dem festgelegten Thema zu haben. Außerdem sollten zusätzlich Ein- sowie Ausschlusskriterien festgelegt werden (Vom Brocke et al. 2009). Nach vom Brocke et al. (2009) ist das Concept Mapping eine sinnvolle Möglichkeit, um Schlüsselkonzepte zu identifizieren.

„Literaturrecherche“ stellt die dritte Phase der strukturierten Literaturrecherche dar. Dieser Schritt umfasst eine Datenbank-, Schlagwort-, Rückwärts- und Vorwärtssuche sowie eine laufende Quellenauswertung (Vom Brocke et al. 2009). Vom Brocke et al. (2009) empfiehlt sich in diesem Schritt auf Artikel aus wissenschaftlichen Zeitschriften oder Tagungsbände von renommierten Konferenzen zu fokussieren. Die Stichworte werden in die zuvor bestimmten Datenbanken eingeben (Vom Brocke et al. 2009). Mit verschiedenen Suchoperatoren kann die Anzahl der Resultate verringert werden, bis eine angemessene Anzahl an Suchergebnissen vorliegt. Die verwendeten Suchbegriffe sollten genau dokumentiert werden, damit Andere beurteilen können, ob sie ausreichend zu der Forschungsfrage passen (Vom Brocke et al. 2009). Literatur, die in den Artikeln zitiert wird, sollte ebenfalls geprüft werden. Außerdem sollten zusätzliche Quellen betrachtet werden, welche den gefundenen Artikel zitiert haben (Vom Brocke

et al. 2009). Nach der Sammlung von genügend Quellen müssen diese in der vierten Phase „Literaturanalyse“ analysiert und synthetisiert werden. Die gefundenen Quellen müssen in dieser Phase inhaltlich ausgewertet werden. Dabei müssen die Titel, Abstracts und unter Umständen auch Volltexte betrachtet werden (Vom Brocke et al. 2009). Dies gelingt mit einer Konzeptmatrix. In dieser werden themenbezogene Konzepte in verschiedene Analyseeinheiten unterteilt. Das Arrangieren, Diskutieren und Synthetisieren von früheren Forschungsergebnissen werden dadurch ermöglicht (Vom Brocke et al. 2009). In der abschließenden Phase „Forschungsagenda“ soll eine Forschungsagenda erstellt werden. Diese dokumentiert die Ergebnisse und fasst diese zusammen (Vom Brocke et al. 2009). Diese Forschungsagenda kann auf Basis der Konzeptmatrix weiterentwickelt werden (Vom Brocke et al. 2009). Der Rechercheumfang der strukturierten Literaturrecherche lässt sich mit der Anwendung der Taxonomie von Cooper (1988) wie folgt festlegen:

Der Fokus dieser Arbeit liegt auf der Vorstellung verschiedener Praktiken zur Behandlung von Ausreißern in der Datenvorverarbeitung für das Data Mining. Das Ziel der strukturierten Literaturrecherche ist es, einen Überblick über bestehende Behandlungsverfahren von Ausreißern in der Datenvorverarbeitung für das Data Mining zu geben und zentrale Aspekte zu identifizieren. Dadurch, dass die Literaturrecherche methodisch aufgebaut ist, ist es möglich die Behandlungsverfahren nach einem ähnlich aufgebauten Konzept vorzustellen. Dadurch soll es ermöglicht werden, die Verfahren im anschließenden Kapitel einzuordnen. Die Ergebnisse werden neutral dargestellt und sollen allgemeine Wissenschaftler ansprechen. Außerdem wird eine repräsentative Menge an Quellen ausgewählt. In der Phase der Konzepterstellung wurden die Datenbanken zur Suche der Quellen definiert. Diese sind die Datenbanken Scopus, ACM Digital Library, Springer Link, ScienceDirect, IEEE Xplore Digital Library und Wiley Online Library. Um grundlegendes Wissen zu erlangen wurden Begriffe, welche sich aus der Forschungsfrage ableiten lassen, in die Datenbanken eingegeben und Teile der Ergebnisse in Kapitel 2 vorgestellt. Daraus konnten weitere Schlüsselbegriffe wie Punktausreißer (point outlier), Klassifikation (classification), überwachtes Lernen (supervised learning) und Clusteranalyse (clustering) abgeleitet werden. Es wurden jeweils die deutschen und englischen Schlüsselbegriffe in die Datenbanken eingegeben. Des Weiteren wurden in dieser Literaturrecherche nur Publikationen betrachtet, die zwischen den Jahren 1996 und 2023 veröffentlicht wurden, da in dem Jahr 1996 das CRISP-DM zum ersten Mal in der Literatur erwähnt wurde. Außerdem wurden Quellen genutzt, die auf die Praktik der Behandlungsverfahren eingehen und diese erläutern. Aus Quellen, bei denen auch die Anwendung einer Ausreißerbehandlung beschrieben wurde, wurden nur die Teile für die strukturierte Literaturrecherche berücksichtigt, die die Praktik erläutern. Es wurden pro Suche die ersten 100 Suchergebnisse betrachtet.

Im Folgenden sind die Schlüsselbegriffe zur besseren Übersicht in Tabelle 2 zusammengefasst, wobei jeweils der deutsche und englische Schlüsselbegriff aufgelistet ist.

Tabelle 2 Übersicht der abgeleiteten Suchbegriffe

Schlüsselbegriffe für die strukturierte Literaturrecherche
Wissensentdeckung in Datenbanken, knowledge discovery in databases
Data Mining
Datenvorverarbeitung, data preprocessing
Ausreißer, outlier
Lokaler Ausreißer, local outlier
Globaler Ausreißer, global outlier
Punktausreißer, point outlier
Kontextualer Ausreißer, contextual outlier
Kollektiver Ausreißer, collective outlier
Klassifikation, classification
Regression
Überwachtes Lernen, supervised learning
Unüberwachtes Lernen, unsupervised learning
Clusteranalyse, clustering
Winsorisieren, winsorizing
Behandlung, treatment
Umgang, handling
Technik, technique
Methode, method

Die Schlüsselbegriffe aus Tabelle 2 wurden einzeln und in verschiedenen Kombinationen mit Hilfe von dem Operator AND in die ausgewählten Suchmaschinen eingegeben. Die Anzahl an gefundenen Ergebnissen wurde in einer Übersicht festgehalten. In Anhang A ist ein beispielhafter Teil einer Suche in der Datenbank Scopus abgebildet. Diese Übersicht befindet sich in Tabelle A-1. Einige Quellen konnten mit unterschiedlichen Schlüsselbegriffen mehrfach gefunden werden. Dazu zählen unter anderem die Literaturwerke von Runkler (2020) und Aggarwal (2017). Im Anschluss an die Vorauswahl der Quellen wurde die Rückwärts- und Vorwärtssuche durchgeführt. Insgesamt konnten 41 Quellen ermittelt werden. Eine Gesamtübersicht befindet sich ebenfalls in Anhang A in Tabelle A-2. In den nachfolgenden Abschnitten werden die Ergebnisse der strukturierten Literaturrecherche vorgestellt. Auf Grund der teilweise nicht ausreichend hohen Informationsdichte bezüglich der Eignung zur Ausreißerbehandlung werden nicht alle Verfahren vorgestellt, die mit Hilfe der strukturierten Literaturrecherche ermittelt werden konnten. Zu Beginn der folgenden Abschnitte wird eine Übersicht über die Quellen dargestellt, welche mit Hilfe der strukturierten Literaturrecherche ermittelt werden konnten. Im Anschluss daran werden die Verfahren beschrieben und die Autoren diskutiert.

3.1 Clusteranalyse

Tabelle 3 Quellenübersicht für die Clusteranalyseverfahren

Autor (Jahr)	Titel der Quelle	Behandlung von Ausreißern	Verfahren
Aggarwal (2015)	Data Mining	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren
Aggarwal (2017)	Outlier Analysis	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren
Aguinis et al. (2013)	Best-Practice Recommendations for Defining, Identifying and Handling Outliers	Ersetzen, Entfernen, Einfluss Verringern	Clusterverfahren, Regressionsverfahren, Trimmen und Winsorisieren
Alasadi und Bhaya (2017)	Review of Data Preprocessing Techniques	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren
Andreopoulos et al. (2009)	A Roadmap of Clustering Algorithms	Entfernen	Clusterverfahren
Bacher (2010)	Clusteranalyse	Entfernen	Clusterverfahren

Fortsetzung Tabelle 3

Barbará (2002)	Requirements for Clustering Data Streams	Entfernen	Clusterverfahren
Chen und Wang (2008)	The Data Mining Technology Based on CIMS and its Application on Automotive Remanufacturing	Entfernen	Clusterverfahren
Cleve und Lämmel (2020)	Data Mining	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren, Klassifikation
D'Onofrio und Meier (2021)	Big Data Analytics	Entfernen	Clusterverfahren
Ester und Sander (2000)	Knowledge Discovery in Databases	Entfernen	Clusterverfahren, Klassifikation
Gama (2010)	Knowledge Discovery from Datastreams	Ersetzen	Clusterverfahren
Hotho (2004)	Clustern mit Hintergrundwissen	Entfernen	Clusterverfahren
Jannaschk (2017)	Infrastruktur für ein Data Mining Design Framework	Entfernen	Clusterverfahren

Fortsetzung Tabelle 3

Mallikharjuna et al. (2023)	Data Preprocessing Techniques	Entfernen	Clusterverfahren
Olson und Lauhoff (2023)	Deskriptives Data-Mining	Entfernen	Clusterverfahren
Petersohn (2005)	Data Mining	Entfernen	Clusterverfahren
Runkler (2020)	Data Analytics	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren, Klassifikation
Sharafi (2012)	Knowledge Discovery in Databases	Entfernen	Clusterverfahren, Klassifikation

In Tabelle 3 ist zu sehen, dass einige der gefundenen Quellen auf die Clusteranalyse zur Behandlung von Ausreißern verweisen. Im Allgemeinen werden bei der Clusteranalyse Daten in verschiedene Segmente gruppiert (D'Onofrio und Meier 2021). Das Ziel liegt darin, dass Daten innerhalb eines Clusters möglichst ähnlich zueinander sind und sich die einzelnen Cluster untereinander unterscheiden (Sharafi 2012). Nach D'Onofrio und Meier (2021) ist ein wichtiger Schritt die Festlegung der Ähnlichkeit der Datensätze. Das gilt für alle Clusteranalyseverfahren. Bei diesen Verfahren werden Ausreißer identifiziert, um sie im Anschluss entfernen zu können. Mit Hilfe der Clusteranalyse werden hauptsächlich Punktausreißer sowie kontextuale Ausreißer behandelt (Chandola et al. 2009). Außerdem gehören Clusterverfahren zu den unüberwachten Verfahren.

Die Integration einer Distanzdimension hilft dabei, die Gruppen der Datenobjekte und ihre Ähnlich- bzw. Unähnlichkeit identifizieren zu können (Sharafi 2012). Die Bestimmung der Unähnlichkeit wird mit Hilfe der Attributausprägungen der einzelnen Datenobjekte durchgeführt (D'Onofrio und Meier 2021). Hierbei wird zwischen numerischen und nominalen Attributen unterschieden. Das Distanzmaß für numerische Attribute schließt aus dem absoluten Abstand zwischen den Objekten auf die Unähnlichkeit (D'Onofrio und Meier 2021). Nach D'Onofrio und Meier (2021) gibt es drei Distanzmaße für numerische Attribute, die als am gebräuchlichsten bezeichnet werden können.

Es existieren zwei Datensätze x und y mit den jeweiligen Attributausprägungen $x = (x_1, x_2, \dots, x_n)$ und $y = (y_1, y_2, \dots, y_n)$ aus denen sich die folgenden Berechnungsformeln für numerische Attribute ergeben:

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

Die euklidische Distanz wird oft bei numerischen Attributen verwendet, da sich für jeden Datensatz ein Punkt im n-dimensionalen Raum definieren lässt (D'Onofrio und Meier 2021).

$$\text{Manhatten-Distanz: } d\,(x,y) = |x_1 - y_1| + \cdots + |x_n - y_n|$$

Die Manhatten-Distanz stellt die Summe der absoluten Differenzwerte dar und ist weniger anfällig für Ausreißer (D'Onofrio und Meier 2021).

$$\text{Maximums-Metrik: } d\,(x,y) = \max\left(|x_1 - y_1|, \dots, |x_n - y_n|\right)$$

Die Distanzberechnung der Maximums-Metrik erfolgt als größter Differenzwert über alle Merkmale (D'Onofrio und Meier 2021).

In der Praxis enthalten die Datensätze häufig nominale und numerische Attribute (D'Onofrio und Meier 2021). In diesem Fall kann der Gower-Koeffizient verwendet werden, da dieser beide Attribute berücksichtigt sowie eine Normierung vornimmt (D'Onofrio und Meier 2021).

$$\text{Gower-Koeffizient: } d\,(x,y) = \frac{1}{n}\sum_{i=1}^{n} d^i\,(x,y)$$

$$\text{Mit: } d^i\,(x,y) = \frac{|x_i - y_i|}{R_i}$$

für numerische Attribute mit R_i = Spannweite (größter Wert – kleinster Wert des i-ten Attributs)

$$\text{Und: } d^i\,(x,y) = \begin{cases} 1, falls\ x_i \neq y_i \\ 0, falls\ x_i = y_i \end{cases}$$

für nominale Attribute.

Zusätzlich zu der Distanzermittlung muss die Aufteilung der Datenobjekte in verschiedene Cluster betrachtet werden (D'Onofrio und Meier 2021). Hierzu werden multivariate statistische Verfahren verwendet (Petersohn 2005). Die Aufteilung kann grundsätzlich in zwei Verfahren

vorgenommen werden, welche als hierarchische Verfahren und partitionierende Verfahren beschrieben werden (D'Onofrio und Meier 2021).

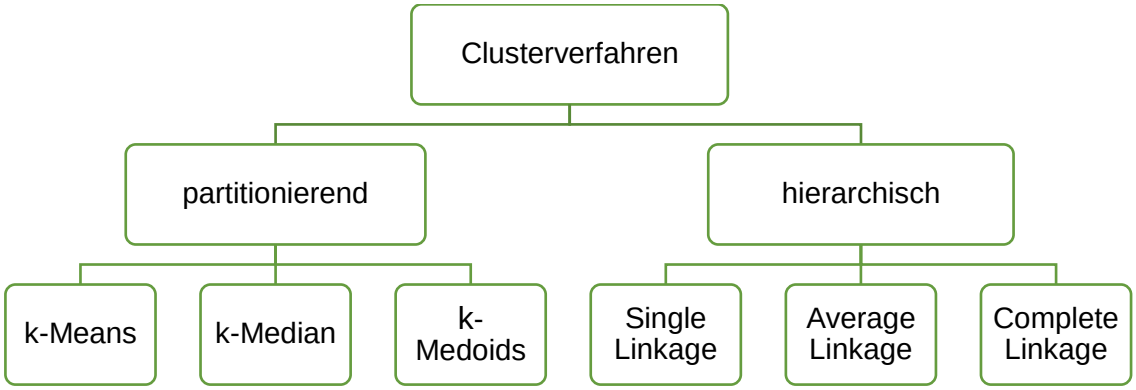

Abbildung 10 Übersicht Clusterverfahren

In Abbildung 10 sind die in dieser Arbeit vorgestellten Verfahren übersichtlich in partitionierende und hierarchische Verfahren aufgeteilt. Es existieren neben den partitionierenden und hierarchischen Clusteranalysen auch die dichtebasiere Clusterbildung sowie das Clustern mit neuronalen Netzen, welche auf Grund des Umfangs dieser Arbeit im Folgenden nicht vorgestellt werden. Hierarchische Verfahren lassen sich in zwei verschiedene Vorgehensweisen unterteilen (Petersohn 2005). Divisive Verfahren gehen von einem alles beinhaltenden Cluster für alle Datensätze aus, welche schrittweise aufgeteilt werden müssen (Petersohn 2005). Die agglomerativen Verfahren gehen entgegengesetzt vor. Alle Datensätze werden jeweils als eigene Cluster definiert, welche dann schrittweise in größere Cluster zusammengefasst werden (Petersohn 2005). Bei diesem Vorgehen müssen mindestens zwei Cluster gebildet werden (Sharafi 2012). Nach D'Onofrio und Meier (2021) sind die agglomerativen Verfahren deutlich relevanter in der Praxis, weshalb sie im Folgenden weiter beschrieben werden.

Zunächst werden zwei Cluster mit der geringsten Distanz zu einem Cluster zusammengefasst (D'Onofrio und Meier 2021). Um diese geringste Distanz zu ermitteln wird eine Distanzmatrix genutzt, in welcher die Distanzen zwischen allen Clustern abgebildet werden (D'Onofrio und Meier 2021). Im Anschluss daran werden mit Hilfe verschiedener Vorgehensweisen weitere Entfernungen zwischen den bereits zusammengefassten Clustern berechnet (D'Onofrio und Meier 2021). Diese Vorgehensweisen können zum Beispiel die sogenannten Single Linkage, Average Linkage und Complete Linkage Verfahren sein. Wird die Minimaldistanz zwischen zwei enthaltenen Objekten zur Bestimmung des Abstandes zwischen zwei Gruppen verwendet, handelt es sich um Single Linkage (D'Onofrio und Meier 2021). Der Average Linkage nutzt dafür den mittleren Abstand der Objektpaare, zum Beispiel mit Hilfe der bereits beschriebenen Manhatten-Distanz (D'Onofrio und Meier 2021). Bei dem Complete Linkage Verfahren wird der Maximalabstand von zwei Objekten berechnet (D'Onofrio und Meier 2021). Alle drei der genannten Verfahren sind für kleine Datenmengen geeignet (Hotho 2004). Außerdem werden sie bei Datensätzen angewandt, welche aus numerischen Attributen bestehen (Hotho 2004). Sobald ein neues Cluster ermittelt wurde, ist eine neue Distanzmatrix zu erstellen (D'Onofrio und Meier 2021). Diese Verfahren lassen sind relativ leicht anwenden (Petersohn 2005). Ein Nachteil von hierarchischen Verfahren ist, dass einmal getroffene Zuordnungen in nachfolgenden Schritten nicht mehr zu korrigieren sind (Petersohn 2005). Da diese Vorgehensweise außerdem bei einer großen Menge an Datenobjekten sehr zeitintensiv ist, werden in der Praxis oft partitionierende Verfahren gewählt, welche weniger exakt als hierarchische Verfahren sind, aber für eine große Datenmenge viel Zeitersparnis mit sich bringen (D'Onofrio und Meier 2021).

Bei der Eignung von hierarchischen Verfahren zur Behandlung von Ausreißern ist sich die gefundene Literatur uneinig. Petersohn (2005) argumentiert, dass sich auf Grund der Tatsache, dass einmal falsch getroffene Zuordnungen im nachfolgenden Schritt nicht mehr korrigierbar sind, hierarchische Verfahren nur bedingt eignen um Ausreißer zu erkennen und zu behandeln. Andreopoulos et al. (2009) sehen das Complete Linkage Verfahren als eher empfindlich gegenüber Ausreißern, da Ausreißer bei dieser Vorgehensweise viel Einfluss haben. Dagegen ist das Average Linkage Verfahren besser zur Ausreißerbehandlung geeignet, denn bei diesem Verfahren können Ausreißer keinen besonderen Einfluss ausüben (Andreopoulos et al. 2009). Nach Bacher (2010) kann auch das Single Linkage Verfahren zur Analyse von Ausreißern genutzt werden. Dies liegt daran, dass das Single Linkage „von einer zu ‚schwachen' Vorstellung der Homogenität in den Clustern ausgeht" (Bacher 2010, S. 152). Werte können bei diesem Verfahren als Ausreißer identifiziert werden, da sie erst in den letzten Schritten der Verschmelzung mit anderen Clustern verknüpft werden.

Das Ziel der partitionierenden Verfahren besteht darin, möglichst optimale Partitionen zu finden (Sharafi 2012). Außerdem wird im Vorfeld die Anzahl der Cluster festgelegt (Gama 2010). Bei diesen Verfahren werden die Cluster durch den jeweiligen Schwerpunkt, auch Centroid genannt, vertreten (Cleve und Lämmel 2020). Daran anschließend werden dann die verbleibenden Objekte schrittweise zugeordnet (D'Onofrio und Meier 2021). Um die Clusterschwerpunkte zu bestimmen gibt es verschiedene Möglichkeiten. Im einfachsten Fall erfolgt eine Selektion der ersten n Datensätze als erste Clusterschwerpunkte (D'Onofrio und Meier 2021). Eine andere Möglichkeit zur partionierenden Clusteranalyse ist das iterative k-Means Verfahren (D'Onofrio und Meier 2021). Dieses Verfahren ist im Allgemeinen effizient in der Verarbeitung von großen Datenmengen (Sharafi 2012). Es eignet sich bei Datensätzen, welche aus Textdaten bestehen (Hotho 2004). Der erste Schritt in diesem Verfahren besteht darin, intiale Cluster zu generieren (Cleve und Lämmel 2020). Diese können zum Beispiel mit Hilfe eines Zufallsgenerators erzeugt werden (Cleve und Lämmel 2020). Im Anschluss daran werden die Clusterschwerpunkte, die sogenannten Centroiden, jedes Clusters berechnet. Diese werden im zweidimensionalen Raum über den jeweiligen Mittelwert der x- und y-Koordinaten ermittelt (Cleve und Lämmel 2020). Nun können Punkte einem neuen Cluster zugeordnet werden (Cleve und Lämmel 2020). Hier wird zur Berechnung der Abstände aller Punkte zu den Centroiden, also den Schwerpunkten, die oben bereits beschriebene Euklidische Distanz verwendet, was auch bedeutet, dass nur numerische Attribute verwendet werden können (D'Onofrio und Meier 2021). Bei dieser Vorgehensweise werden schrittweise für alle über bleibenden Datensätze die Distanzen zu den Clusterschwerpunkten berechnet und im Anschluss die Datensätze dem Centroid zugeordnet, welcher den geringsten Abstand hat (D'Onofrio und Meier 2021). Der neue Centroid dieses Clusters wird daraufhin erneut aus den Mittelwerten der einzelnen Merkmalsausprägungen der enthaltenen Datensätze berechnet (D'Onofrio und Meier 2021). Dies wird so lange gemacht, bis keine Verbesserung der Zuordnung mehr erreicht werden kann, also kein Datenpunkt mehr sein Cluster wechselt (Cleve und Lämmel 2020). In der folgenden Abbildung 10 ist eine beispielhafte Clusterbildung dargestellt.

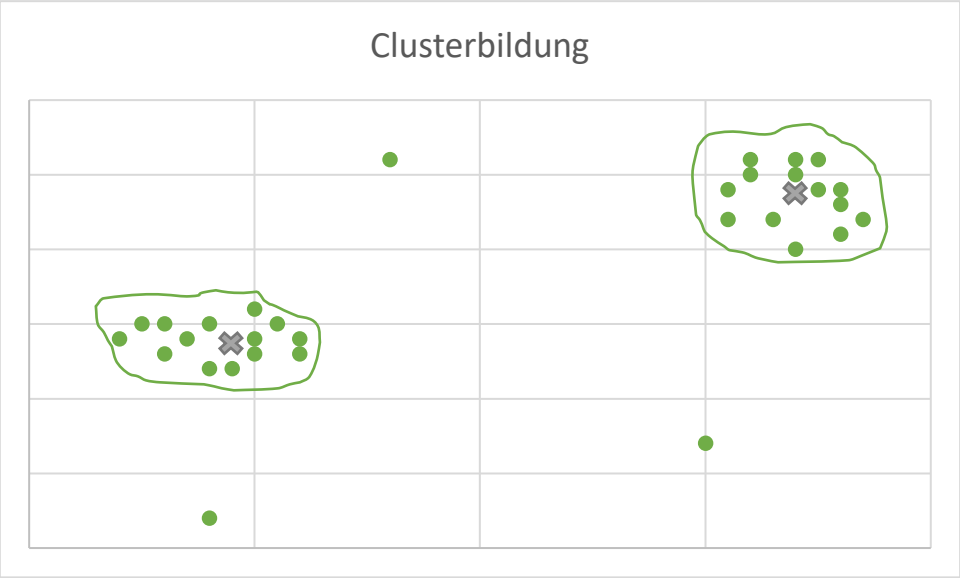

Abbildung 11 Clusterbildung in Anlehnung an Alasadi und Bhaya (2017), S. 4

Die Abbildung 11 zeigt, wie eine Clusterbildung aussehen kann. Es ist deutlich zu erkennen, dass zwei Cluster und drei Ausreißer in dem Datensatz vorliegen. Die Ausreißer sind dadurch zu erkennen, dass sie keinem der Cluster eindeutig zuzuordnen sind. Ein Vorteil dieses Verfahrens ist, dass es oft nach einer geringen Anzahl an Iterationen eine stabile Verteilung der Datenpunkte erreicht (Cleve und Lämmel 2020). Allerdings ist es sensibel im Hinblick auf die erste Wahl der Cluster (Sharafi 2012). Des Weiteren kann die k-Means Methode empfindlich gegenüber Ausreißern sein. Diese ziehen den Clusterschwerpunkt in ihre Richtung und somit besteht die Gefahr, dass ein Cluster verzerrt wird (Cleve und Lämmel 2020).

Ein weiteres Verfahren innerhalb der Clusteranalyse ist k-Medoids. Dieses ist eng mit dem k-Means verwandt (Sharafi 2012). Es hat ebenso zum Ziel, die Daten in Cluster mit einem minimalen quadrierten Abstand der einzelnen Datenpunkte zu partitionieren (Sharafi 2012). Allerdings wird bei dem k-Medoids Verfahren anstatt eines Centroiden ein Medoid verwendet (Sharafi 2012). Ein Medoid zeichnet sich dadurch aus, dass er die kleinste durchschnittliche Entfernung zu den anderen Punkten im Cluster hat (Sharafi 2012). Diese Entfernung kann mit dem Gower-Koeffizienten berechnet werden. Der Medoid ist außerdem Teil der Datenmenge (Aggarwal 2015). Es besteht die Möglichkeit, die Qualität der Clusterbildung durch Tauschen zu verbessern (Cleve und Lämmel 2020). In jedem Schritt kann ein Nichtmedoid mit einem Medoiden den aktuellen Status tauschen (Cleve und Lämmel 2020). Dadurch, dass der Schwerpunkt nicht mehr das zentrale Element eines Clusters ist, sondern ein Datenpunkt ausgewählt wird, kann das Cluster durch einen Ausreißer nicht verzerrt werden und es ist weniger empfindlich gegenüber Ausreißern als das k-Means Verfahren (Cleve und Lämmel 2020). Anzumerken ist an dieser Stelle, dass in der Literatur Ausreißer häufig mit Rauschen gleichgesetzt werden. So argumentiert zum Beispiel Sharafi (2012), dass das k-Medoids-Verfahren auf Grund der Medoide weniger von Ausreißern beeinflusst werden kann und das Verfahren auf Grund dessen robuster hinsichtlich Rauschen ist.

Eine dritte Möglichkeit zur Behandlung von Ausreißern ist das k-Median Verfahren. Hier wird anstelle des Centroiden oder Medoiden der Median verwendet (Cleve und Lämmel 2020). Bei diesem Verfahren wird eine Mischung aus den beiden vorher beschriebenen Verfahren benutzt. In jeder Komponente wird der Median berechnet (Cleve und Lämmel 2020). Der Punkt muss also kein Teil der bestehenden Datenmenge sein, allerdings muss jede Komponente einen realen Wert haben (Cleve und Lämmel 2020). Das k-Median Verfahren arbeitet mit der

Manhattan-Distanz (Aggarwal 2015). Nach Aggarwal (2015) ist dieses Verfahren robuster als k-Means, da der Median weniger empfindlich gegenüber Ausreißern ist als der Mittelwert.

Allgemein haben Clustermethoden den Vorteil, dass sie im Vergleich zu abstandsbasierten Methoden schneller sind (Aggarwal 2017). Allerdings können sie bei kleineren Datenbeständen unter Umständen nicht ausreichend tiefe Einblicke in die vorliegenden Daten geben (Aggarwal 2017). Nach Aggarwal (2015) können Ausreißer dazu neigen, in kleinen eigenen Clustern aufzutreten. Dies liegt daran, dass Anomalien in der Datenerhebung unter Umständen einige Male wiederholt wurden (Aggarwal 2015). Laut Olson und Lauhoff (2023) hat das k-Means Clusterverfahren den Vorteil, dass mögliche Datenfehler wie zum Beispiel Rauschen und Ausreißer eliminiert werden. Sie haben allerdings festgestellt, dass es helfen kann den Mittelwert mit dem Median zu ersetzen, da dieser von Ausreißern nicht beeinflusst wird (Olson und Lauhoff 2023). Andere Autoren, wie zum Beispiel Aggarwal (2015), treffen nicht die Aussage, dass der Median vollständig unbeeinflusst von Ausreißern ist. Sie sehen den Median allerdings auch als besser geeignet an, da er zwar nicht vollständig unbeeinflusst von Ausreißern ist, allerdings wesentlich weniger als der Mittelwert. Aggarwal (2015) merkt an, dass das k-Means Verfahren zu suboptimalen Ergebnissen neigt, wenn ein Ausreißer als anfänglicher Repräsentant für das Verfahren ausgewählt wird. In diesen Fällen kann es vorkommen, dass der Repräsentant eines Clusters in einer leeren Region ist, also für das Cluster nicht repräsentativ (Aggarwal 2015). Bei der Betrachtung der Aussagen, dass das k-Means Verfahren Schwächen in Bezug auf Ausreißer aufzeigen, ist sich die gefundene Literatur einig. Auch das k-Medoids Verfahren kann laut Aggarwal (2015) eine robuste Alternative zu dem k-Means Verfahren sein. In der Regel wird ein Ausreißer bei der Verwendung des k-Medoid Verfahrens während eines iterativen Austausches verworfen. Allerdings stagniert das Verfahren, wenn in einem nachfolgenden Iterationsschritt zum Beispiel ein leeres Cluster auftritt. Laut Aggarwal (2015) kann dieses Problem behoben werden, wenn zusätzlich ein Schritt eingefügt wird, der sehr kleine Cluster verwirft und diese durch zufällig ausgewählte Punkte aus den Daten ersetzt werden. Es fällt auf, dass einige Autoren die Begriffe Ausreißer und Rauschen nicht klar voneinander trennen, wie zum Beispiel Sharafi (2012). Er bewertet das k-Medoids Verfahren als besser geeignet bei Ausreißern als das k-Means Verfahren und argumentiert, dass Medoide weniger stark von Ausreißern beeinflusst werden und dass das k-Medoids Verfahren deshalb robuster gegen Rauschen ist. Die Autoren der Literaturwerke sind sich einig, dass unter den partionierenden Verfahren das k-Means Verfahren am empfindlichsten auf Ausreißer reagiert. Sowohl die Vorgehensweise des k-Medoids und des k-Median sind robuster gegenüber Ausreißern. Zusätzlich empfehlen D'Onofrio und Meier (2021) bezüglich der Wahl eines Distanzmaßes die Manhatten-Distanz bei der Anwendung von Clusterverfahren zu nutzen, da sich diese als weniger empfindlich gegenüber Ausreißern herausgestellt hat. Daraus kann abgeleitet werden, dass das k-Means Verfahren weniger zur Ausreißerbehandlung geeignet ist, da hier die Euklidische Distanz verwendet wird. Das k-Median Verfahren dagegen verwendet die Manhatten-Distanz, was schließen lässt, dass es besser für die Behandlung von Ausreißern geeignet ist. Anhand der gefundenen Aussagen in den Literaturwerken lässt sich weiterhin erkennen, dass bei der Anwendung eines Clusterverfahrens meist auf ein partitionierendes Verfahren zurückgegriffen wird, wenn es um die Behandlung von Ausreißern geht. Trotzdem können auch mit hierarchischen Verfahren Ausreißer gefunden und anschließend entfernt werden. Aus der strukturierten Literaturrecherche geht hervor, dass die partitionierenden Verfahren besser für große Datenmengen verwendet werden sollten und die hierarchischen Verfahren für kleinere Datenmengen. Die meisten vorgestellten Verfahren können nur Datensätze mit ordinalen oder metrischen Skalenniveaus behandeln.

3.2 Regression

Tabelle 4 Quellenübersicht für die Regressionsverfahren

Autor (Jahr)	Titel	Behandlung von Ausreißern	Verfahren
Aggarwal (2015)	Data Mining	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren
Aggarwal (2017)	Outlier Analysis	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren
Agostinelli et al. (2015)	Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination	Ersetzen	Regressionsverfahren
Aguinis et al. (2013)	Best-Practice Recommendations for Defining, Identifying and Handling Outliers	Ersetzen, Entfernen, Einfluss Verringern	Clusterverfahren, Regressionsverfahren, Trimmen und Winsorisieren
Alasadi und Bhaya (2017)	Review of Data Preprocessing Techniques	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren
Alpaydin (2019)	Maschinelles Lernen	Ersetzen, Entfernen	Regressionsverfahren, Klassifikation

Fortsetzung Tabelle 4

Barnett (2004)	Environmental statistics	Ersetzen, Entfernen, Einfluss verringern	Regressionsverfahren, Trimmen und Winsorisieren
Cleve und Lämmel (2020)	Data Mining	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren, Klassifikation
Cohen et al. (2003)	Applied Multiple Regression/Correlation Analysis for the Behavioral Sience	Ersetzen	Regressionsverfahren
Ghavami (2020)	Big data analytics methods	Ersetzen	Regressionsverfahren
Giloni et al. (2006)	Robust weighted LAD regression	Ersetzen	Regressionsverfahren
Han et al. (2012)	Data Mining. Concepts and Techniques	Ersetzen	Regressionsverfahren
Khan et al. (2007)	Robust Linear Model Selection Based on Least Angle Regression	Ersetzen, Entfernen, Einfluss verringern	Regressionsverfahren, Trimmen und Winsorisieren
Runkler (2020)	Data Analytics	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren, Klassifikation

Fortsetzung Tabelle 4

Yuan und Bentler (1998)	Structural Equation Modeling with Robust Covariances	Ersetzen	Regressionsverfahren
Zhong und Yuan (2011)	Bias and Efficiency in Structural Equation Modeling	Ersetzen	Regressionsverfahren

Wie bereits in Abschnitt 2.5 angemerkt, können Ausreißer mit Verfahren der Regression behandelt werden. Dies geht ebenfalls aus dem Auszug der Quellen der strukturierten Literaturrecherche in Tabelle 4 hervor. Verfahren der Regression zählen zu den überwachten Verfahren. Außerdem werden sie oft bei der Behandlung von Datensätzen mit numerischen Skalenniveaus eingesetzt (Han et al. 2012; Cleve und Lämmel 2020). Im Allgemeinen werden Regressionsmethoden verwendet, um aus unabhängigen Variablen, auch als Prädikatorvariablen bezeichnet, abhängige Variablen, auch als Antwortvariablen bezeichnet, vorherzusagen (Han et al. 2012). Ausreißer können mittels der Regression behandelt werden, indem sie durch die Werte der Regressionsfunktion ersetzt werden (Cleve und Lämmel 2020). Allgemein beschreibt die Regressionsfunktion einen Zusammenhang zwischen numerischen Attributen (Cleve und Lämmel 2020). Sie schätzt die Art der Beziehung zwischen den Merkmalen (Runkler 2020). In der Literatur werden die x-Werte oft als Inputvariablen und die y-Werte als Outputvariablen bezeichnet (Aggarwal 2015). Im Folgenden werden einige Regressionsverfahren vorgestellt. Nach Cleve und Lämmel (2020) geht die lineare Regression der kleinsten Quadrate (engl. ordinary least squares, OLS) davon aus, dass die Beziehung zwischen x- und y-Werten durch eine Gerade beschrieben werden kann. Mit ihrer Hilfe kann eine Aussage darüber getroffen werden welche Werte Merkmale annehmen müssen, um die geforderte Qualität zu erreichen (Runkler 2020). Sei b der Regressionskoeffizient und a der y-Achsenabschnitt, so lässt sich diese Beziehung mit der folgenden Regressionsfunktion beschreiben:

$$y_i \approx b * x_i + a$$

Das Ziel der linearen Regression ist es, die Fehler zwischen dem errechneten Wert und dem tatsächlichen Zielwert zu minimieren (Cleve und Lämmel 2020).

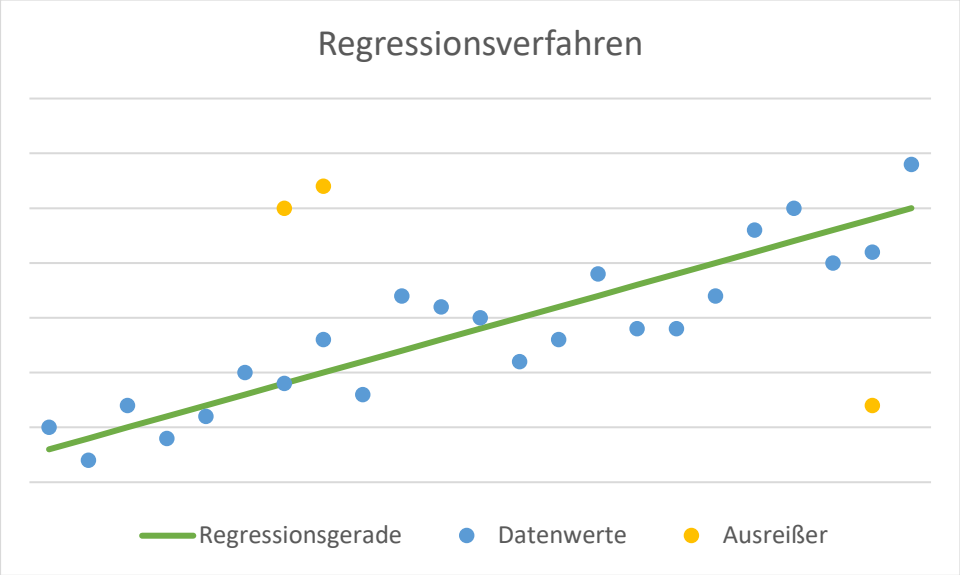

Abbildung 12 graphische Darstellung der Regressionsgleichung nach Ghavami (2020), S. 105

In Abbildung 12 ist graphisch eine Regressionsgleichung dargestellt. Die gelben Punkte stellen die Ausreißer dar, welche deutlich von der Regressionsgeraden abweichen. Um keine Kompensation zwischen negativen und positiven Abweichung zu erhalten, wird bei dieser Herangehensweise der quadratische Fehler untersucht (Cleve und Lämmel 2020). Der quadratische Regressionsfehler E lässt sich nach Runkler (2020) wie folgt formulieren:

$$E = \frac{1}{n}\sum_{k=1}^{n}(y_k - b * x_k - a)^2$$

Dieser Wert beschreibt die Abweichung eines Punktes von der Regressionsgeraden, also seinem vorhergesagten Wert (Ghavami 2020). Ein Indikator dafür, wie gut das Regressionsmodell zu den Daten passt, ist R^2, auch bezeichnet als das Bestimmtheitsmaß (Ghavami 2020). Dieser Wert befindet sich zwischen 0 und 1. Wenn das Modell sehr gut zu den tatsächlichen Werten passt ist $R^2 = 1$. Passen das Regressionsmodell und die tatsächlichen Werte sehr schlecht zueinander ist $R^2 = 0$ (Ghavami 2020). Da Ausreißer den Fehler beeinträchtigen können, ist die lineare Regression empfindlich gegenüber Ausreißern (Runkler 2020). Die robuste Regression dagegen schränkt den Einfluss von Ausreißern ein (Runkler 2020).

Ein alternatives Vorgehen in der Regression ist die kleinste absolute Abweichung (engl. least absolute deviation, LAD). Hier werden die Werte des Regressionskoeffizienten b so gewählt, dass die Summe der absoluten Werte minimiert wird (Cohen et al. 2003). Da die Differenz nicht quadriert wird, ist dieses Verfahren bei hohen Abweichungen besser geeignet als die lineare Regression (Cohen et al. 2003). Trotzdem ist das LAD Verfahren nicht robust gegenüber ungewöhnlichen Prädikatorvariablen, er hat also einen niedrigen Breakdown-Punkt (Giloni et al. 2006). Aus diesem Grund besteht die Möglichkeit, das LAD Verfahren zu gewichten um damit eine höhere Robustheit zu erlangen (Giloni et al. 2006). Sei w das Gewicht, so lässt sich nach Giloni et al. (2006) das Verfahren der gewichteten kleinsten absoluten Abweichung (WLAD) wie folgt beschreiben:

$$\sum_{i=1}^{n} w_i |y_i - x_i * b| \rightarrow min$$

Das WLAD Verfahren kann als LAD-Regression mit entsprechend transformierten Daten behandelt werden (Giloni et al. 2006). Außerdem kann der WLAD durch die geringere quadratische Verzerrung besser mit Ausreißern umgehen als andere Verfahren (Giloni et al. 2006).

Ein weiteres Beispiel für die robuste Regression ist die Summe der kleinsten quadratischen Fehler (engl. Least Trimmed Squares, LTS). Bei diesem Verfahren werden die quadratischen Regressionsfehler für jeden der n Fälle vom niedrigstem zum höchsten sortiert (Cohen et al. 2003). Im Anschluss daran wird eine beliebige Grenze festgelegt und nur die m kleinsten Fehler die unter dieser Grenze liegen werden berücksichtigt (Cohen et al. 2003). Der Anteil der zu vernachlässigenden quadratischen Regressionsfehler kann zum Beispiel bei 20% liegen. Beim LTS ist anzumerken, dass es im Allgemeinen gut funktioniert, allerdings auf eine Anhäufung von Ausreißern immer noch empfindlich reagieren kann (Cohen et al. 2003).

Eine andere Möglichkeit der robusten linearen Regression ist die M-Schätzung. Hier werden die kleinsten quadratischen Regressionsfehler gewichtet (Cohen et al. 2003). Das Gewicht w richtet sich danach, wie weit der berechnete Fehler von der Regressionsgraden abweicht (Cohen et al. 2003). Die Fehler, die auf der Geraden liegen oder sehr nah sind, gehen mit der vollen Gewichtung ein, also $w = 1$ (Cohen et al. 2003). Je weiter der Fehler von der Regressionsgeraden abweicht, desto geringer wird er gewichtet (Cohen et al. 2003).

Die Empfindlichkeit der quadratischen Fehlerfunktion in der linearen Regression gegenüber Ausreißern ist laut Runkler (2020) darin begründet, dass die Ausreißer einen quadratischen Einfluss auf den Fehler haben. Auch Ghavami (2020) geht darauf ein, dass Ausreißer einen großen Einfluss auf die lineare Regressionsgrade haben können, da diese von der Summe der Quadrate abhängig ist und nicht von der Summe der einfachen Abstände. In einigen Quellen, wie zum Beispiel bei Ghavami (2020), werden Ausreißer auch als Anomalien bezeichnet. Han et al. (2012) stellen fest, dass in einigen Data Mining Verfahren Ausreißer als Rauschen verworfen werden. Auch Alasadi (2017) trennt die Begriffe Ausreißer und Rauschen nicht. Daher kann seine Aussage, dass Regression zur Glättung von verrauschten Daten geeignet ist, auch auf die Ausreißer bezogen werden. Cohen et al. (2003) führen aus, dass der LTS im Allgemeinen gut funktioniert. Sie merken allerdings an, dass der LTS in sehr seltenen Fällen eine ungenaue Schätzung erzeugen kann, wenn eine Ansammlung von Ausreißern im Datenbestand vorhanden ist. Ab wann eine Menge an Datenwerten als Ansammlung gilt, wird nicht

beschrieben. Auch der LAD kann potentiell eine deutlich verbesserte Schätzung erzeugen als der OLS wenn Ausreißer in den Daten vorhanden sind. Dies liegt laut Cohen et al. (2003) daran, dass bei diesem Verfahren die Differenz nicht quadriert wird. Außerdem ist der M-Schätzer in Fällen mit großen Abweichungen signifikant robuster als die lineare Regression (Cohen et al. 2003; Han et al. 2012). Aus der strukturierten Literaturrecherche geht hervor, dass die lineare Regression als eher empfindlich gegenüber Ausreißern einzuordnen ist. Ebenso kann festgehalten werden, dass durch das Verwenden von Verfahren der robusten Regression der quadratische Einfluss von Ausreißern eingeschränkt wird, da diese keine Quadrierung verwenden. Die Aussagen der Autoren der Literaturwerke stimmen überein, dass bei der Behandlung von Ausreißern robuste Regressionsverfahren einer linearen Regression vorgezogen werden sollten. Ghavami (2020) merkt außerdem an, dass eine angemessene Anzahl an Datenpunkten verfügbar sein muss, um ein gutes Regressionsmodell erstellen zu können. Werden zu viele Datenpunkte ausgewählt, wird das Modell instabil und lässt sich wahrscheinlich nicht wiederholen (Ghavami 2020). Ghavami (2020) geht nicht darauf ein, was als angemessene Anzahl an Datenpunkten gilt.

Da die Regressionsverfahren zu den überwachten Verfahren zählen, kann zusätzlich mit Abschnitt 2.5 abgeleitet werden, dass sie sich zur Behandlung von Punktausreißern und kollektiven Ausreißern eignen. Außerdem können die vorgestellten Regressionsverfahren hauptsächlich Datensätze analysieren, welche aus numerischen Attributen bestehen.

3.3 Klassifikation

Tabelle 5 Quellenübersicht für die Klassifikationsverfahren

Autor (Jahr)	Titel	Behandlung von Ausreißern	Verfahren
Alpaydin (2019)	Maschinelles Lernen	Ersetzen, Entfernen	Regressionsverfahren, Klassifikation
Bramer (2016)	Principles of Data Mining	Entfernen	Klassifikation
Cleve und Lämmel (2020)	Data Mining	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren, Klassifikation
Ester und Sander (2000)	Knowledge Discovery in Databases	Entfernen	Clusterverfahren, Klassifikation
Lee et al. (2020)	Automatic Bridge Design Parameter Extraction for Scan-to-BIM	Entfernen	Klassifikation
Luengo et al. (2020)	Big Data Preprocessing	Entfernen	Klassifikation

Fortsetzung Tabelle 5

Radovanovic et al. (2015)	Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection	Entfernen	Klassifikation
Runkler (2020)	Data Analytics	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren, Klassifikation
Sharafi (2012)	Knowledge Discovery in Databases	Entfernen	Clusterverfahren, Klassifikation

Mit Hilfe der strukturierten Literaturrecherche konnte festgestellt werden, dass sich ebenfalls Verfahren der Klassifikation für die Ausreißerbehandlung eignen. Die gefundenen Quellen sind in Tabelle 5 abgebildet. Klassifikationsverfahren behandeln Ausreißer auf die gleiche Art wie die Clusteranalyseverfahren. Nachdem die vorhandenen Ausreißer identifiziert wurden, werden sie entfernt. Des Weiteren kann die Klassifikation den überwachten Verfahren zugeordnet werden (D'Onofrio und Meier 2021). Bei Klassifikationsverfahren gibt es zwei Vorgehensweisen, zwischen denen unterschieden werden kann (Cleve und Lämmel 2020). Instanzbasierte Verfahren klassifizieren Datensätze mit der Hilfe schon gegebene Beispieldatensätze (Cleve und Lämmel 2020). Diese Verfahren sind einfach gehalten und es wird kein Modell entwickelt (Cleve und Lämmel 2020). Um ein Objekt zu klassifizieren, wird unter anderem das ähnlichste Objekt aus den Beispieldatensätzen gesucht und mit Hilfe dessen die Klasse vorhergesagt (Cleve und Lämmel 2020). Diese instanzbasierten Verfahren können auch als Lazy Learner bezeichnet werden (Cleve und Lämmel 2020). Es gibt andere Verfahren, welche auf Grundlage der gegebenen Beispieldatensätze ein Modell berechnen (Cleve und Lämmel 2020). Hier werden neue Objekte nicht direkt mit Hilfe der Beispieldatensätze klassifiziert, sondern ausschließlich mit Hilfe des berechneten Modells (Cleve und Lämmel 2020). Diese Verfahren werden auch als Eager Learner bezeichnet (Cleve und Lämmel 2020). Im Folgenden wird das k-Nearest-Neighbour-Verfahren (kNN Verfahren) beschrieben, welches zu den instanzbasierten Verfahren zu zählen ist und damit als Lazy Learner bezeichnet werden kann.

Die grundlegende Idee dieses Verfahrens ist die unbekannte Klasse eines Objektes mit Hilfe der Objekte, die dem unbekannten am nächsten sind, schätzen zu können (Bramer 2016). Eine Voraussetzung für die Anwendung dieses Verfahrens ist, dass ein Abstandsmaß für die Daten existiert (Cleve und Lämmel 2020). Des Weiteren können metrische, nominale oder reellwertige Attribute mit diesem Verfahren behandelt werden (Cleve und Lämmel 2020). Nach D'Onofrio und Meier (2021) kann das kNN Verfahren unter anderem bei Bilddaten verwendet werden. Bei diesem Verfahren werden neue Objekte mit Hilfe der errechneten Ähnlichkeit zu bereits gespeicherten Objekten klassifiziert (Cleve und Lämmel 2020). Eine Menge an k Objekten, die zu dem neuen Objekt am ähnlichsten sind, bilden die Grundlage zur Vorhersage der Klasse des neuen Objekts (Runkler 2020). Die bereits vorhandenen Objekte werden auf ihre Klasse untersucht und die, die am häufigsten auftritt, wird als neue Klasse des neuen Objekts vorhergesagt (Alpaydin 2019). Ein weiterer Punkt, der in Betracht gezogen werden muss, ist die Wahl von k. Wird k zu klein gewählt, ist das kNN Verfahren sensibel gegenüber Ausreißern (Ester und Sander 2000). Wird k verhältnismäßig groß gewählt, kommen Objekte aus anderen Klassen mit in die Entscheidungsmenge (Ester und Sander 2000). Nach Ester und Sander (2000) liefert ein mittleres k allgemein eine hohe Klassifikationsgüte, daher sollte für k folgendes gelten: $1 \ll k < 10$

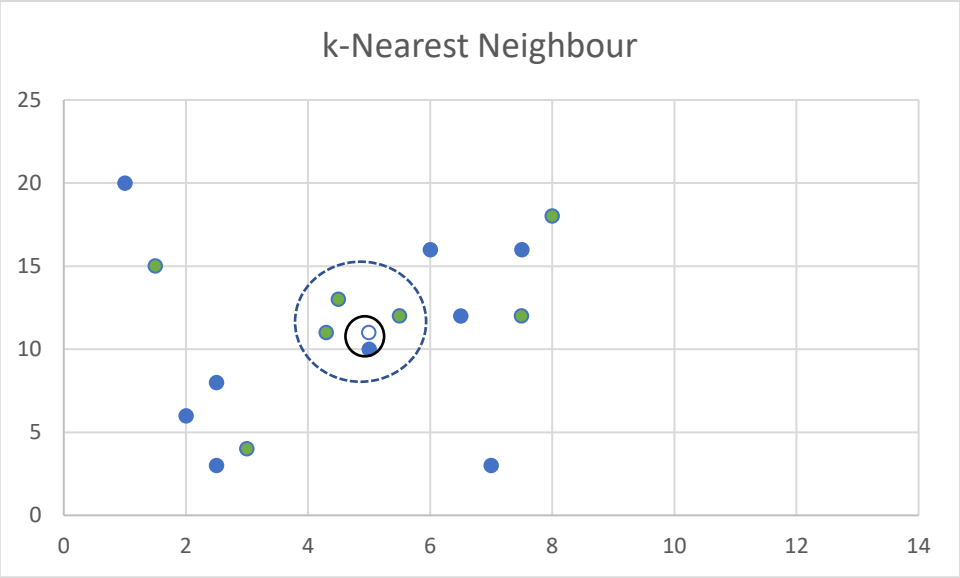

Abbildung 13 kNN Verfahren in Anlehnung an Cleve und Lämmel (2020), S. 88

In Abbildung 13 ist eine Veranschaulichung des kNN Verfahrens abgebildet. Die blauen und grünen Datenpunkte stellen jeweils eine Klasse dar. Die Klasse des weißen Punktes ist die, welche vorhergesagt werden soll. Es ist zu erkennen, dass die Wahl von k ausschlaggebend für das Ergebnis sein kann. Wählt man $k = 1$ wird die Klasse des neuen Objekts als blau vorhergesagt, was an der durchgezogenen Linie zu erkennen ist. Wählt man aber $k = 4$ ist die Vorhersage der Klasse eine andere, und zwar grün. Dies ist durch die gestrichelte Linie ersichtlich. Nach Cleve und Lämmel (2020) werden in der Praxis mehrere Varianten mit verschiedenen k berechnet, um die jeweils vorhergesagten Klassen vergleichen zu können und eine abschließende Entscheidung treffen zu können. Die Berechnung der Klasse mit metrischen Attributen wird von Cleve und Lämmel (2020) beschrieben und erfolgt wie folgt:

Seien $V = \{v_1, v_2, \dots, v_m\}$ die endliche Menge der Werte, die das Zielattribut annehmen kann, $f: \mathbb{R}^n \to V$ die diskrete Funktion, x_i die Trainingsbeispiele und y das zu klassifizierende Objekt. Dann lautet die Klassifizierungsfunktion:

$$klasse\ (y) = \max \sum_{p=1}^{k} \delta\left(v, f\left(x_p\right)\right)$$

$$\text{mit } \delta(a,b) = \begin{cases} 1, falls\ a = b \\ 0, sonst \end{cases}$$

Liegen die Daten als nominale Daten vor, kann der Abstand mit der Hamming-Distanz berechnet werden (Cleve und Lämmel 2020). Die Formel für die Hamming-Distanz lautet wie folgt:

$$dist_H(x,y) = count_i(x_i \neq y_i)$$

Dieses Distanzmaß zählt, an wie vielen Punkten sich die gegebenen Datensätze unterscheiden (Cleve und Lämmel 2020). Allerdings erkennt es nicht, wie groß der Unterschied ist. Liegen reellwertige Zielattribute vor, so sei die Funktion gegeben durch $f: \mathbb{R}^n \to \mathbb{R}$. Der Unterschied zu einem diskreten Fall besteht darin, dass der Mittelwert der Zielattributwerte zurückgegeben wird (Cleve und Lämmel 2020). Die Formel dafür lautet wie folgt:

$$f'(y) = \frac{\sum_{p=1}^{k} f\left(x_p\right)}{k}$$

Anschließend an die Klassifizierungsfunktion wird ein Schwellenwert auf der Grundlage der Standardabweichung berechnet (Lee et al. 2020). Ein Problem des kNN Verfahrens kann darin gesehen werden, dass sich die Ähnlichkeit von zwei Objekten an ihrem Abstand bemessen

lässt (Cleve und Lämmel 2020). Um die Größe des Abstandes mit in das Verfahren einzubeziehen, kann der Abstand gewichtet werden (Cleve und Lämmel 2020). Das Vorgehen kann nach Cleve und Lämmel (2020) als Shepard's method beschrieben werden. Die Formel für einen diskreten Datensatz lautet wie folgt:

$$klasse(y) = \begin{cases} f(x_i) & falls\ y = x_i\ \text{für}\ ein\ i \\ max \sum_{p=1}^{k} w_p * \delta\left(v, f(x_p)\right) & sonst \end{cases}$$

$$\text{Mit: } \delta(a,b) = \begin{cases} 1, falls\ a = b \\ 0, sonst \end{cases} \qquad \text{und: } w_p = \frac{1}{dist(y,x_p)^2}$$

Das Gewicht wird in dieser Methode mit der Inversen des Quadrats der Distanz zwischen dem Trainingsdatensatz und dem neuen Datensatz beschrieben (Cleve und Lämmel 2020).

Auch für reellwertige Funktionen besteht die Möglichkeit, die Shepard's method anzuwenden.

Die Zielattributwerte für y ergeben sich mit $w_p = \frac{1}{dist(y,x_p)^2}$ wie folgt:

$$f'(y) = \begin{cases} f(x_i)\ falls\ y = x_i\ \text{für}\ ein\ i \\ \frac{\sum_{p=1}^{k} w_p * f(x_p)}{\sum_{p=1}^{k} w_p} \quad sonst \end{cases}$$

Nach Cleve und Lämmel (2020) bieten sich die gleichen Verfahren zur Behandlung von Ausreißern an wie bei der Behandlung von fehlenden Werten. Dazu zählt auch die Klassifikation. Radovanovic et al. (2015) definieren den umgekehrten kNN-Zähler als Ausreißerschwellenwert. Überschreitet ein Datenpunkt den vom Anwender festgelegten Schwellenwert, so kann dieser als Ausreißer definiert werden (Radovanovic et al. 2015). Lee et al. (2020) definieren die Ausreißer auch als Rauschen, wie bereits bei anderen Autoren aufgefallen ist. Sie beziehen sich auf das gleiche Argument wie Radovanovic et al. (2015). Werte, die einen Schwellenwert überschreiten, können als Ausreißer erkannt werden und im Anschluss entfernt werden. Diese Aussage wird von Lee et al. (2020) allerdings auf die Fälle beschränkt, in denen das Rauschen niedrig genug ist. Überschreitet das Rauschen einen bestimmten Wert, werden Ausreißer nicht mehr erkannt und können in Folge dessen auch nicht mehr behandelt werden. Auf die Größe des Grenzwertes gehen Lee et al. (2020) nicht ein. Daraus ist zu schließen, dass ab einer bestimmten Menge an Ausreißern das kNN Verfahren nicht mehr geeignet ist. Wie außerdem zu Beginn des Abschnittes 3.3 erwähnt, gehören Klassifikationsverfahren zu den überwachten Verfahren. In Kombination mit Abschnitt 2.5 lässt sich ableiten, dass mit diesen Verfahren Punktausreißer sowie kollektive Ausreißer behandelt werden können.

3.4 Trimmen und Winsorisieren

Tabelle 6 Quellenübersicht für Verfahren des Trimmens und Winsorisierens

Autor (Jahr)	Titel	Behandlung von Ausreißern	Verfahren
Aggarwal und Sathe (2017)	Outlier Ensembles	Entfernen, Einfluss verringern	Trimmen und Winsorisieren
Aguinis et al. (2013)	Best-Practice Recommendations for Defining, Identifying and Handling Outliers	Ersetzen, Entfernen, Einfluss Verringern	Clusterverfahren, Regressionsverfahren, Trimmen und Winsorisieren
Barnett (2004)	Environmental statistics	Ersetzen, Entfernen, Einfluss verringern	Regressionsverfahren, Trimmen und Winsorisieren
Blaine (2018)	Winsorizing	Entfernen, Einfluss verringern	Trimmen und Winsorisieren
Chambers et al. (2000)	Winsorization for Identifying and Treating Outliers in Business Surveys	Einfluss verringern	Winsorisieren

Fortsetzung Tabelle 6

Gosh und Vogt (2012)	Outliers: An Evaluation of Methodologies	Entfernen, Einfluss verringern	Trimmen und Winsorisieren
Hoo et al. (2002)	A method of robust multivariate outlier replacement	Einfluss verringern	Winsorisieren
Huber und Ronchetti (2009)	Robust Statistics	Entfernen, Einfluss verringern	Trimmen und Winsorisieren
Khan et al. (2007)	Robust Linear Model Selection Based on Least Angle Regression	Ersetzen, Entfernen, Einfluss verringern	Regressionsverfahren, Trimmen und Winsorisieren
Sullivan et al. (2021)	So Many Ways To Assess Outliers	Entfernen, Einfluss verringern	Trimmen und Winsorisieren
Yi (2023)	Robust and Multivariate Statistical Methods	Entfernen, Einfluss verringern	Trimmen und Winsorisieren

Durch die strukturierte Literaturrecherche konnten weitere Verfahren zur Behandlung von Ausreißern ermittelt werden. Zu diesen können das Trimmen und das Winsorisieren gezählt werden. In Tabelle 6 sind die Quellen abgebildet, die zu diesen Verfahren ermittelt werden konnten. Im ersten Teil dieses Abschnitts wird das Trimmen kurz im Allgemeinen erläutert sowie die Variante des Schwellenwert-Trimmens. Im Anschluss daran wird das Winsorisieren beschrieben.

Das Prinzip des Trimmens ist, einen beliebigen Bereich für Werte festzulegen und alle Werte, die darüber oder darunter liegen, zu eliminieren (Aguinis et al. 2013). Das setzt voraus, dass die Daten in eine Reihenfolge gebracht werden können. Das Skalenniveau muss dementsprechend beim Trimmen aus metrischen Attributen bestehen (Aggarwal und Sathe 2017). Beim Trimmen wird die gleiche Methode zur Behandlung von Ausreißern angewandt wie bereits bei der Clusteranalyse und bei der Klassifikation genannt. Nachdem der Wertebereich festgelegt wurde, werden die außerhalb liegenden Werte, also die Ausreißer, entfernt. Eine Variante davon ist das Schwellenwert-Trimmen (AVG-T) (Aggarwal und Sathe 2017). Bei diesem Verfahren wird nicht wie bei dem zuvor beschriebenen Trimmen ein Prozentsatz verwendet, sondern die Detektoren werden auf der Grundlage ihres Abstands zu den Ensemble-Scores bewertet (Aggarwal und Sathe 2017). Diejenigen, welche einen besonders großen Abstand haben, werden anschließend getrimmt (Aggarwal und Sathe 2017). Um festzustellen, wann Detektoren als abnormal bezeichnet werden können, werden statistische Signifikanztests verwendet (Aggarwal und Sathe 2017). Sei die gesamte Menge der Detektoren m. Nach Aggarwal und Sathe (2017) werden dann die folgenden fünf Schritte zur Berechnung durchgeführt. Zuerst werden die Ergebnisse jedes Detektors auf einen Mittelwert von 0 und eine Standardabweichung standardisiert (Aggarwal und Sathe 2017). Dabei sei der standardisierte Wert des i-ten Punktes für den j-ten Detektor $O(i,j)$. Nun wird der durchschnittliche Wert für

$$a_i = \sum_{j=1}^{m} \frac{O(i,j)}{m}$$

berechnet. Im Anschluss daran wird die aggregierte Detektorabweichung mit

$$\Delta j = \sum_{i=1}^{n} |O(i,j) - a_i|$$

für den j-ten Detektor berechnet (Aggarwal und Sathe 2017). Im vierten Schritt werden alle $\Delta 1, \dots, \Delta m$ auf den Mittelwert 0 und die Standardabweichung standardisiert (Aggarwal und Sathe 2017). Im abschließenden Schritt wird das Trimmen durchgeführt. Es wird jeder Detektor ausgeschlossen, für den $\Delta j > 1$ gilt.Nun wird der bereinigte Mittelwert a'_i für alle übrig gebliebenen Werte berechnet (Aggarwal und Sathe 2017). Mit dieser Methode können ungewöhnliche Werte ausgeschlossen werden, die die Ergebnisse stark beeinflussen würden (Aggarwal und Sathe 2017). Zu diesem Verfahren lässt sich anmerken, dass die Basisdetektoren für alle Punkte getrimmt werden (Aggarwal und Sathe 2017).

Nach Aggarwal und Sathe (2017) kann mit Hilfe des Trimmens der Einfluss von Ausreißern auf die Datenbestände verringert werden. Auch nach Barnett (2004) können robuste Verfahren, zu denen auch das Trimmen zählt, dabei helfen, den Einfluss von Verunreinigungen zu minimieren. Als robuste Schätzung ist das Verfahren des Trimmens intuitiv attraktiv, allerdings ist es nach der Ansicht von Barnett (2004) nicht weit verbreitet. Huber und Ronchetti (2009) merken an, dass relativ hohen Trimmraten (25% oder mehr) eine negative Auswirkung auf die Ausreißerbehandlung haben, insbesondere, wenn die Größe der Stichprobe relativ klein ist. Bei einem Stichprobenumfang von 20 könnte der 10% getrimmte Mittelwert nur mit einem Ausreißer umgehen. Daraus lässt sich schließen, dass das Trimmen eher für große Datenmengen zu empfehlen ist. Gosh und Vogt (2012) bezeichnen das Entfernen eines Ausreißers aus einer Stichprobe als eine extreme Art, den Einfluss eines Ausreißers zu verringern. Daher sollte das Trimmen nur angewandt werden, wenn im Vorfeld geprüft wurde, dass der Ausreißer kein legitimer Wert ist (Gosh und Vogt 2012). Ist dies jedoch gegeben, ist das Verfahren des

Trimmens eine Möglichkeit, Ausreißer zu behandeln. Bezüglich des Trimmens lässt sich festhalten, dass es in seinen Grundzügen zur Ausreißerbehandlung geeignet ist, allerdings einige Beschränkungen beachtet werden sollten.

Das zweite Verfahren, welches in diesem Abschnitt vorgestellt wird, ist das des Winsorisierens. Dabei wird der Einfluss von Ausreißern verringert. Außerdem kann das Winsorisieren den unüberwachten Verfahren zugeordnet werden. Bei diesem Verfahren werden die Datenwerte nicht wie beim Trimmen eliminiert, sondern alle Werte, die außerhalb eines beliebig festgelegten Bereichs liegen, durch weniger extreme Werte ersetzt (Aguinis et al. 2013). Dadurch wird beim Winsorisieren der Einfluss von Ausreißern auf den Mittelwert und die Varianz abgeschwächt (Blaine 2018). Zum Beispiel würden bei einer 90% Winsorisierung alle Werte unterhalb des 5% Quantil auf das 5% Quantil angehoben werden und alle Werte oberhalb des 95% Quantils auf das 95% Quantil herabgesetzt werden (Aguinis et al. 2013). Diese Grenze ist als gängig zu betrachten und wird demnach häufig angewandt (Gosh und Vogt 2012; Blaine 2018). Der daraus resultierende Wert kann als winsorisiert bezeichnet werden (Blaine 2018). Nach Blaine (2018) ist dies ein Verfahren, welches sich neben dem Trimmen zur Ausreißerbehandlung eignet. Seien y_i die beobachteten Werte, $\hat{y}_i$ die angepassten Beobachtungen und die Residuen $r_i = y_i - \hat{y}_i$. Sei außerdem s_i der Schätzwert für den Standardfehler von y_i oder r_i (Huber und Ronchetti 2009). Dann kann das Winsorisieren für die Pseudo-Beobachtung y_i^* nach Huber und Ronchetti (2009) mathematisch wie folgt beschrieben werden:

$$y_i^* = \begin{cases} y_i \;\; falls\; |r_i| \leq cs_i \\ \hat{y}_i - cs_i \;\; falls\; r_i < -cs_i \\ \hat{y}_i + cs_i \;\; falls\; r_i > cs_i \end{cases}$$

Die Konstante c hat in diesem Fall Einfluss auf die Robustheit (Huber und Ronchetti 2009; Hoo et al. 2002). Huber und Ronchetti (2009) nennen als eine gute Wahl für c einen Wert zwischen 1 und 2. Die neuen Pseudo-Beobachtungen werden nun anstelle der ursprünglichen Beobachtungen verwendet, um neue Werte für $\hat{y}_i$ und $r_i = y_i - \hat{y}_i$ sowie s_i zu berechnen (Huber und Ronchetti 2009). Das Vorgehen kann wiederholt werden und dadurch werden die Werte iterativ ersetzt, damit sie näher an den anderen Beobachtungen sind (Hoo et al. 2002). Durch das Winsorisieren der Werte und deren anschließender Ersetzung durch geeignetere Werte ist es möglich, Ausreißer zu erkennen und zu behandeln (Hoo et al. 2002). Nach Hoo et al. (2002) ist es möglich, dass das Vorhandensein von Ausreißern korrekte Werte als anormal erscheinen lassen kann, sodass diese Werte während des Winsorisierens auch ersetzt werden könnten. Dies ist allerdings selten und tritt auf, wenn eine sehr große Menge von Ausreißern in dem Datensatz vorhanden ist (Hoo et al. 2002). Wenn alle Beobachtungen gleich genau sind, dann lautet die Formel für die Schätzung der Varianz s^2 einer Beobachtung nach Huber und Ronchetti (2009) wie folgt:

$$s^2 = \frac{1}{n-p}\sum r_i^2 \qquad \text{mit } n \text{ Beobachtungen und } p \text{ Parametern}$$

Verwendet man nun anstelle von r_i die modifizierten Residuen $r_i^* = y_i^* - \hat{y}_i$ kann die dadurch entstehende Verzerrung korrigiert werden, indem man für die Schätzung der Varianz s^2 folgende Formel verwendet:

$$s^2 = \frac{\frac{1}{n-p}\sum {r_i^*}^2}{(\frac{m}{n})^2} \qquad \text{mit } m \text{ unveränderten Beobachtungen}$$

Nach Hoo et al. (2002) können durch das Winsorisieren der Punkte und dem anschließenden Ersetzen durch geeignete Werte Ausreißer gut erkannt und behandelt werden. Nach Huber und Ronchetti (2009) ist es sogar offensichtlich, dass das Winsorisieren den Einfluss von Ausreißern abschwächt. Der Datensatz wird bereinigt, indem die Ausreißer iterativ angepasst werden, bis Konvergenz erreicht ist. Gosh und Vogt (2012) vertreten die Ansicht, dass das Behandlungsverfahren des Winsorisierens einen guten Kompromiss in der Ausreißerbehandlung

darstellt. Die Gefahr der Verzerrung durch Ausreißer wird dadurch verringert, dass ein angepasster Wert beibehalten wird (Gosh und Vogt 2012). Blaine (2018) verwendet in seinem Literaturwerk ein Beispiel, bei dem die Stichprobe einen Umfang von 10 hat. Er wendet eine 20%-Winsorisierung an. Das Ergebnis kann nach Blaine (2018) als ein besserer Schätzer für die Lage und Variabilität der ursprünglichen Variablen bezeichnet werden. Auf Grundlage dieser Annahme lässt sich schließen, dass das Winsorisieren auf für kleine Datensätze geeignet ist. Chambers et al. (2000) bezeichnen das Winsorisieren als ein robustes Verfahren mit einer bestimmten Einflussfunktion. Einen Beleg dafür, dass mehr Werte nach oben als nach unten korrigiert werden müssen, bringen Chambers et al. (2000) nicht vor. Da die Werte beim Winsorisieren, wie auch beim Trimmen, in eine Reihenfolge gebracht werden müssen, lässt sich schlussfolgern, dass ein metrisches Skalenniveau vorliegen muss. Außerdem kann mit Kapitel 2.5 abgeleitet werden, dass das Winsorisieren Punktausreißer sowie kontextuale Ausreißer behandeln kann, da es zu den unüberwachten Verfahren gehört. Insgesamt lässt sich feststellen, dass die durch die strukturierte Literaturrecherche gefundene Anzahl an Quellen, welche das Vorgehen des Trimmens und des Winsorisierens beschreiben, gering ist. Dies lässt darauf schließen, dass die anderen vorgestellten Verfahren dieses Kapitels in der Praxis eine häufigere Anwendung finden. Trotzdem stellen das Trimmen und das Winsorisieren mögliche Verfahren zur Behandlung von Ausreißern in der Datenvorverarbeitung für das Data Mining dar und wurden aus diesem Grund in dieser Arbeit mit einbezogen.

Durch das Vorstellen und Diskutieren einzelner Verfahren in Bezug auf die Ausreißerbehandlung wurde das zweite Teilziel erreicht. Mit Hilfe des in Kapitel 2 beschriebenen Stand der Technik und der in diesem Kapitel vorgestellten Verfahren werden im anschließenden Kapitel Anforderungen an die Verfahren abgeleitet. Durch die anschließende Auswahl von Kriterien soll eine Methode zur Entscheidungsunterstützung für die Ausreißerbehandlung entwickelt werden, um das Hauptziel zu erreichen.

4 Methode zur Entscheidungsunterstützung zur Behandlung von Ausreißern

In diesem Kapitel wird eine Methode zur Entscheidungsunterstützung zur Behandlung von Ausreißern erarbeitet. Diese soll die Auswahl der möglichen Verfahren einschränken und die Entscheidung für ein Verfahren zur Ausreißerbehandlung vereinfachen. Zunächst werden Anforderungen der in Kapitel 3 mittels der strukturierten Literaturrecherche vorgestellten Verfahren abgeleitet. Mit Hilfe der erwähnten Anforderungen werden Kriterien ausgewählt, die in die Methode zur Entscheidungsunterstützung einfließen. In Abschnitt 4.3 wird die Methode zur Entscheidungsunterstützung in Form eines Entscheidungsbaums entwickelt. Die Wahl der Methode der Entscheidungsunterstützung wird erläutert und die Vorgehensweise erklärt. Im Anschluss an die Entwicklung der Methode der Entscheidungsunterstützung werden die Ergebnisse dieser Arbeit zusammengefasst und diskutiert.

4.1 Ableitung von Anforderungen an die vorgestellten Verfahren

Um eine Methode zur Entscheidungsunterstützung bezüglich der Auswahl eines Behandlungsverfahrens entwickeln zu können, müssen zunächst die Anforderungen an das zu erstellende Modell aus den Erkenntnissen des 3. Kapitels abgeleitet werden. Es wurde deutlich, dass die vorgestellten Verfahren verschiedene Bedingungen voraussetzen, wie zum Beispiel unterschiedliche Datenmengen oder Skalenniveaus. Das Ziel dieses Abschnittes ist es, die Anforderungen der vorgestellten Verfahren in Bezug auf die Behandlung von Ausreißern abzuleiten. Durch die strukturierte Literaturrecherche wurde deutlich, dass die Größe der zu behandelnden Datenmengen einen Einfluss auf die Auswahl des Behandlungsverfahrens bezüglich der Ausreißer hat. Die Verfahren können unterschiedliche Mengen an Daten gut verarbeiten. Aus der strukturierten Literaturrecherche geht hervor, dass zum Beispiel das vorgestellte Klassifikationsverfahren kNN eher für kleinere Datensätze geeignet ist. Dagegen kann das Clusterverfahren der k-Medoids besser mit großen Datenmengen umgehen. In Bezug auf die Datenmenge ist in dieser Arbeit die Anzahl der Datensätze gemeint. Durch die Literaturrecherche konnte keine allgemeine Angabe gefunden werden, ab wann ein Datenbestand als groß und wann als klein gilt. Um die Größe einer Datenmenge trotzdem mit einbeziehen zu können, wird im Rahmen dieser Arbeit die Anzahl nicht als feste Zahl definiert, sondern auf A gesetzt. Die systematische Literaturrecherche hat außerdem gezeigt, dass die unterschiedlichen vorgestellten Verfahren zwischen den Skalenniveaus nominal, ordinal und metrisch unterscheiden. Nominalskalierte Daten können zum Beispiel die Postleitzahl oder eine Farbe sein. Sie lassen sich in Kategorien einteilen, aber sie können nicht in eine natürliche Reihenfolge gebracht werden. Ein Verfahren, welches dieses Skalenniveau als Eingangsvariable verarbeiten kann, ist zum Beispiel das Klassifikationsverfahren kNN. Das k-Means Verfahren kann mit nominalen Attributen nicht umgehen, da hier ein Mittelwert berechnet wird. Da jedoch zum Beispiel kein Mittelwert bei Farben existiert, würde das k-Means Verfahren in diesem Fall nicht angewandt werden können. Ordinalskalierte Daten sind zum Beispiel Schulnoten. Diese können in eine Reihenfolge gebracht werden. Die Abstände können nicht ermittelt werden. Ein Verfahren, welches diese Variablen verarbeiten kann, ist zum Beispiel das Clusterverfahren Single Linkage. Die lineare Regression kann nur metrische Daten verarbeiten. Metrische Daten können in eine Reihenfolge gebracht werden und die Abstände können ermittelt werden, wie zum Beispiel die Geschwindigkeit. Es wird deutlich, dass in der Methode zur Entscheidungsunterstützung die verschiedenen Skalenniveaus berücksichtigt werden müssen. Während der Vorstellung der Verfahren zur Behandlung von Ausreißern in Kapitel 3 ist außerdem festgehalten worden, dass die Art des Datensatzes Einfluss auf die Möglichkeit der Behandlung nimmt. Die Daten können neben anderen als Zeitreihen, Textdaten oder Bilddaten vorliegen. In der Methode zur Entscheidungsunterstützung muss die Art des Datensatzes berücksichtigt werden,

was durch die strukturierte Literaturrecherche ersichtlich wurde. Eine weitere Anforderung an die Verfahren zur Behandlung von Ausreißern wurde bereits in Kapitel 2 thematisiert und während der strukturierten Literaturrecherche aufgegriffen. Es gibt verschiedene Arten von Ausreißern, die eine teilweise unterschiedliche Behandlung erforderlich machen. Punkt- ,kontextuale und kollektive Ausreißer können in verschiedenen Datensätzen vorkommen und können unter Anderem von Kontextattributen oder miteinander verbundenen Beobachtungen abhängig sein. Aus diesem Grund sollten die möglichen Arten eines Ausreißers ebenfalls in der Methode zur Entscheidungsunterstützung berücksichtigt werden.

4.2 Auswahl von Kriterien

Die Auswahl von den Kriterien für die Methode zur Entscheidungsunterstützung ergibt sich aus den im vorherigen Abschnitt abgeleiteten Anforderungen. Für eine bessere Übersicht werden die Kriterien im Folgenden tabellarisch dargestellt.

Tabelle 7 Übersicht der Kriterien

Kriterium	Ausprägungen
Datenmenge	klein, groß
Skalenniveau	nominal, ordinal, metrisch
Art des Datensatzes	Zeitreihe, Textdaten, Bilddaten
Art des Ausreißers	Punktausreißer, kontextualer Ausreißer, kollektiver Ausreißer

Aus Tabelle 7 und den in Abschnitt 4.1 abgeleiteten Anforderungen geht hervor, dass ein Kriterium die Datenmenge darstellt. Sie wird im Rahmen dieser Arbeit entweder als $< A$ oder als $\geq A$ bezeichnet. Ein weiteres Kriterium ist das Skalenniveau. Für die Methode zur Entscheidungsunterstützung kann zwischen einer nominalen, ordinalen oder metrischen Skala unterschieden werden. Außerdem wird die Art des Datensatzes berücksichtigt. Im Rahmen dieser Arbeit werden Zeitreihen, Textdaten und Bilddaten berücksichtigt und in die Methode zur Entscheidungsunterstützung mit einbezogen. Als letztes Kriterium wird die Art des Ausreißers benannt. Wie in Kapitel 2 bereits beschrieben, werden Punktausreißer, kontextuale Ausreißer und kollektive Ausreißer betrachtet und in die Methode zur Entscheidungsunterstützung mit einbezogen.

4.3 Entwicklung der Methode zur Entscheidungsunterstützung

Als Methode zur Entscheidungsunterstützung zur Verfahrenswahl bezüglich der Behandlung von Ausreißern wird ein Entscheidungsbaum gewählt. Ein Entscheidungsbaum hat eine hierarchische Struktur, welche aus einer Wurzel, mehreren Zweigen und Knoten sowie Blättern besteht. Die vorgestellten Verfahren aus Kapitel 3 sollen die letzten Blätter des Entscheidungsbaums darstellen. Da es viele Möglichkeiten zur Behandlung von Ausreißern gibt, bietet sich

der Entscheidungsbaum auf Grund seiner Übersichtlichkeit an. Des Weiteren ist ein Entscheidungsbaum beliebig erweiterbar, was im Hinblick auf weitere vorhandene Verfahren in der Literatur zur Ausreißerbehandlung in der Datenvorverarbeitung von Vorteil ist. Mit Hilfe des Entscheidungsbaums soll es ermöglicht werden, in verhältnismäßig kurzer Zeit eine geeignete Entscheidung bezüglich der Auswahl eines Behandlungsverfahrens treffen zu können. Des Weiteren sind durch einen Entscheidungsbaum die wichtigen Kriterien und deren Ausprägungen schnell zu erkennen. Der Entscheidungsbaum soll die im Rahmen dieser Arbeit vorgestellten Verfahren abbilden und einordnen. Es besteht die Möglichkeit, in Zukunft weitere Verfahren zu ergänzen. Der Entscheidungsbaum wird mit Hilfe der in den Abschnitten 4.1 und 4.2 erarbeiteten Anforderungen und Kriterien erstellt. Die Verfahren, welche die letzten Blätter darstellen, sind nicht als einzige Lösung zu betrachten. Es werden nur die in Kapitel 3 erwähnten Anforderungen des jeweiligen Verfahrens berücksichtigt. In Abbildung 14 ist das Grundgerüst abgebildet.

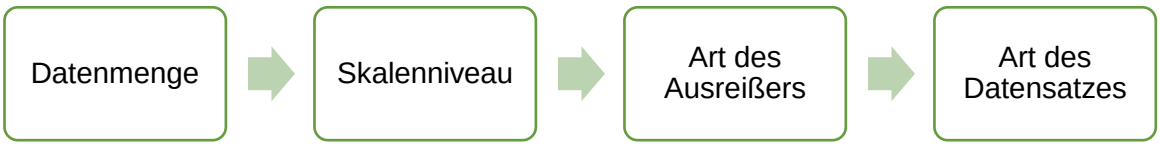

Abbildung 14 Grundgerüst des Entscheidungsbaumes

Das erste Kriterium zur Einordnung eines Datenbestandes stellt die Datenmenge dar. Im Entscheidungsbaum wird zwischen klein und groß unterschieden. Die Datenmenge wird als erstes Kriterium gewählt, da durch sie eine erste Einteilung der Attribute vorgenommen werden kann. Außerdem soll die Methode zur Entscheidungsunterstützung zu Beginn eher eine geringe Anzahl an Zweigen haben, damit die Übersichtlichkeit gegeben bleibt. Die Datenmenge wird mit in den Entscheidungsbaum einbezogen, da es bei der Auswahl eines Behandlungsverfahrens von Ausreißern sehr wichtig ist, ein Verfahren zu wählen, welches mit der vorhandenen Datenmenge umgehen kann. Auch wenn in der Literatur keine einheitlichen Angaben darüber gefunden werden konnten, ab wann eine Datenmenge als groß gilt, geht aus einem großen Teil der Quellen hervor, dass die Menge ausschlaggebend sein kann, ob ein Verfahren zuverlässig Ausreißer finden und sie behandeln kann. Daher wird wie in Abschnitt 4.1 bereits beschrieben, die Menge auf A festgesetzt. Im Entscheidungsbaum wird eine kleine Datenmenge mit $< A$ beschrieben und eine große Datenmenge mit $\geq A$.

Die darauffolgenden Knoten sollen das Kriterium des Skalenniveaus abbilden. Das Skalenniveau ist eine weitere Voraussetzung, welche nicht vernachlässigt werden darf. Die in dieser Arbeit betrachteten Skalenniveaus nominal, ordinal und metrisch haben einen großen Einfluss darauf, welches Verfahren zur Behandlung von Ausreißern ausgewählt werden kann. Von nominalen Attributen, wie zum Beispiel Farben, kann der Abstand nicht mit der euklidischen Distanz berechnet werden. Daher kann bei nominal skalierten Daten kein Verfahren zur Behandlung von Ausreißern angewandt werden, bei welchem nur die euklidischen Distanz verwendet wird. Daher sollte dieses Kriterium im Entscheidungsbaum abgebildet werden, damit die Wahl auf ein Verfahren fällt, welches mit dem Skalenniveau der Eingangsvariablen umgehen kann.

Das Kriterium, welches anschließend an das Skalenniveau betrachtet wird, ist die Art der Ausreißer. Wie aus der strukturierten Literaturrecherche und der Ableitung der Anforderungen hervor geht, werden in dieser Arbeit die Arten Punktausreißer, kontextueller Ausreißer und kollektiver Ausreißer betrachtet. Durch die strukturierte Literaturrecherche wurde ebenfalls deutlich, dass einige Verfahren besser für kontextuelle Ausreißer und andere besser für kollektive

Ausreißer geeignet sind. Daher ist es wichtig, die Art des Ausreißers auch in den Entscheidungsbaum mit einzubeziehen. Dadurch soll vermieden werden, dass der Anwender des Entscheidungsbaumes ein Verfahren auswählen könnte, welches kontextuale oder kollektive Ausreißer nicht erkennen kann. Im Gegensatz zu den kontextualen und kollektiven Ausreißern und wie in Abschnitt 2.5 beschrieben, können die Punktausreißer mit unüberwachten und überwachten Verfahren erkannt werden. Daher werden sie in der Methode zur Entscheidungsunterstützung jeweils einen Knoten mit den kontextualen Ausreißern und einen mit den kollektiven Ausreißern bilden.

Das letzte Kriterium ist die Art des Datensatzes. In dieser Arbeit werden, wie bereits beschrieben, Zeitreihen, Textdaten und Bilddaten berücksichtigt. Es gibt weitere Arten von Datensätzen. Diese werden aber auf Grund nicht ausreichender Berücksichtigung in den ermittelten Literaturwerken aus der Methode zur Entscheidungsunterstützung ausgeschlossen. Die verschiedenen Arten von Datensätzen stehen im Zusammenhang mit dem Skalenniveau und der Art des Ausreißers. Wie in Abschnitt 2.5 beschrieben, ist ein Ausreißer kontextual, wenn er als einzelner Ausreißer in einer Zeitreihe vorliegt. Liegen mehrere Ausreißer in einer Zeitreihe vor, sind diese kollektiv. Aus Gründen der Übersichtlichkeit werden einzelne Zweige zusammengefasst, das sie zu gleichen Ergebnissen kommen. Zum Beispiel empfiehlt der Entscheidungsbaum für kleine Datenmengen, nominal skalierte Daten, Punkt- und kollektive Ausreißer sowie Text- oder Bilddaten das kNN Verfahren. Einzelne Zweige für Text- und Bilddaten sind nicht notwendig, da keine weiteren Verfahren ermittelt wurden, die unter den zuvor genannten Voraussetzungen mit dieser Art von Datensätzen umgehen können. Aus diesen Gründen ist es sinnvoll, die Art der Datensätze als letztes Entscheidungskriterium festzulegen.

Im Folgenden ist die Methode zur Entscheidungsunterstützung in Form eines Entscheidungsbaumes abgebildet. Aus Gründen der Übersichtlichkeit und der einfacheren Lesbarkeit wurde der Entscheidungsbaum auf die folgenden Seiten aufgeteilt. Zunächst ist in Abbildung 15 der Entscheidungsbaum für den Fall zu sehen, dass die Datenmenge $< A$ ist. Im Anschluss daran ist in Abbildung 16 der Entscheidungsbaum für eine Datenmenge $\geq A$ zu sehen. Ebenfalls befindet sich in beiden Abbildungen eine Legende für die verwendeten Abkürzungen.

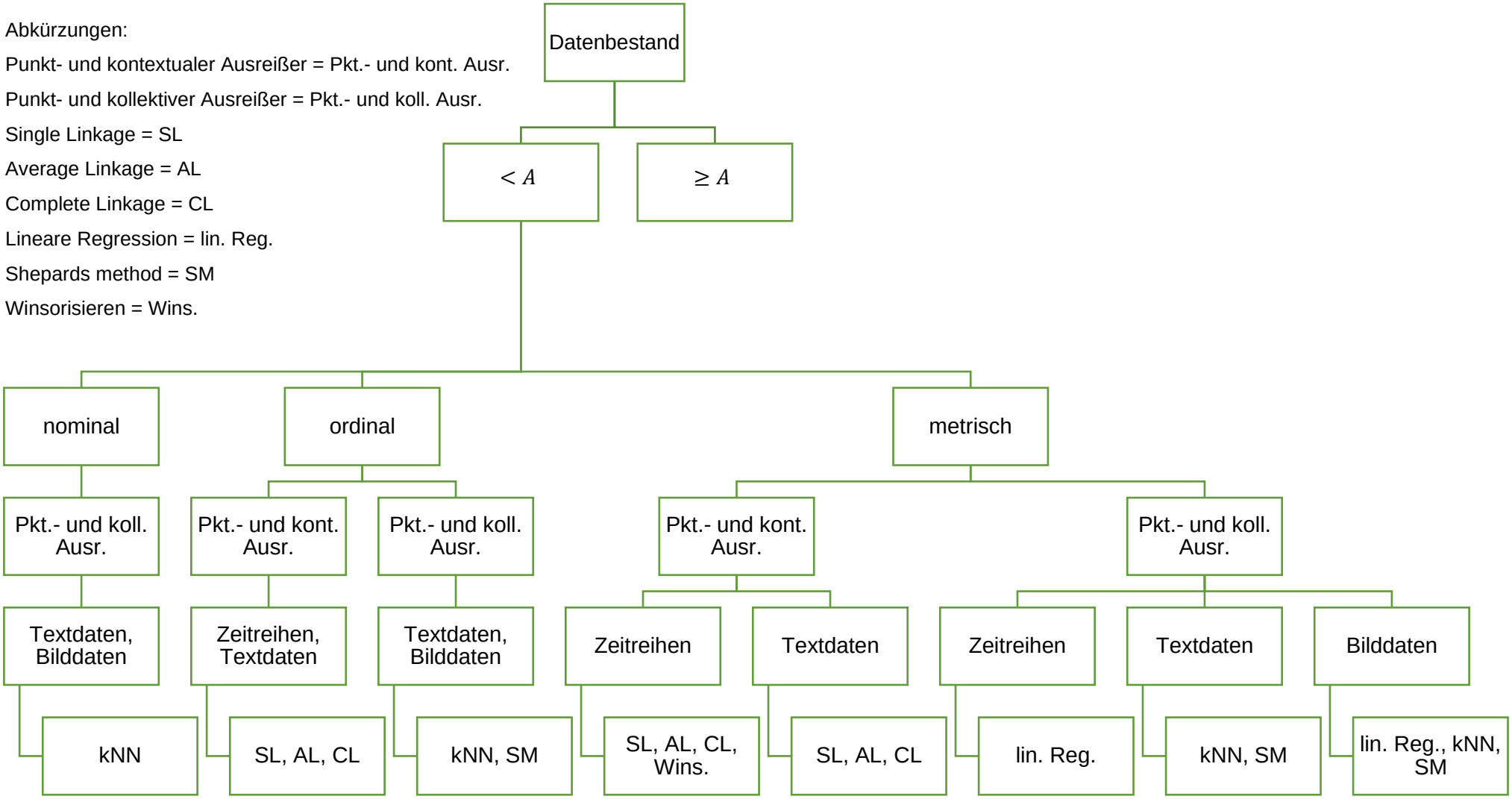

Abbildung 15 Entscheidungsbaum für Datenmengen <A

Abkürzungen:

Punkt- und kontextualer Ausreißer = Pkt.- und kont. Ausr.

Punkt- und kollektiver Ausreißer = Pkt.- und koll. Ausr.

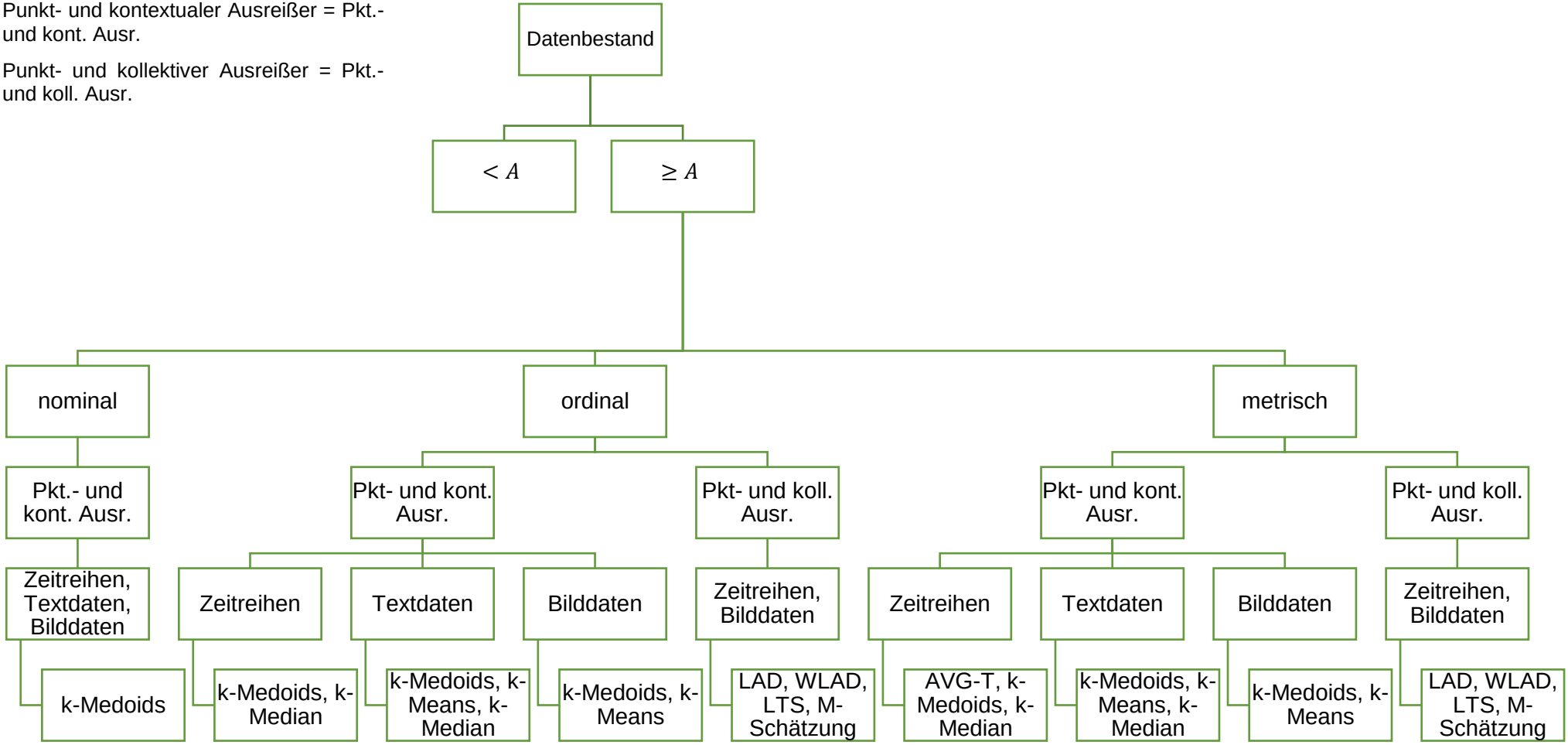

Abbildung 16 Entscheidungsbaum für Datenmengen ≥ A

Im Folgenden wird die Methode zur Entscheidungsunterstützung mit Hilfe eines Fallbeispiels exemplarisch besprochen.

In dem Fallbeispiel handelt es sich um ein Bilderarchiv, welches in Bildersuchmaschinen genutzt wird. Es ist bekannt, dass der Datenbestand aus nominal skalierten Daten besteht. Außerdem sollen in diesem Schritt der Datenvorverarbeitung Punktausreißer behandelt werden, damit anschließend das Data Mining angewandt werden kann. Es ist weiterhin bekannt, dass es sich um einen großen Datenbestand handelt. Der vorliegende Entscheidungsbaum soll die Auswahl des Behandlungsverfahren in Bezug auf Ausreißer unterstützen.

Der vorliegende Datenbestand muss zunächst in die Kriterien des Entscheidungsbaumes eingeordnet werden. Aus dem Fallbeispiel geht hervor, dass es sich um eine große Datenmenge handelt. Daher befinden wir uns in Abbildung 16, in welcher der Entscheidungsbaum für Datenmengen $\geq A$ abgebildet ist. Das anschließende Kriterium ist das Skalenniveau. In dem oben genannten Fallbeispiel handelt es sich um nominal skalierte Daten. Des Weiteren ist bekannt, dass Punktausreißer behandelt werden sollen. Diese treten im Entscheidungsbaum immer in Kombination mit entweder kontextualen oder kollektiven Ausreißern auf. Betrachtet man die möglichen Zweige nach dem Skalenniveau wird deutlich, dass nur Punkt- und kontextuale Ausreißer behandelt werden können. Für Punkt- und kollektive Ausreißer konnte kein Behandlungsverfahren gefunden werden, weshalb dieser Zweig aus dem Entscheidungsbaum ausgeschlossen wurde. Die Datenbestände, die behandelt werden können, sind Zeitreihen, Text- oder Bilddaten. Da in dem Fallbeispiel Bilddaten vorliegen, werden die Anforderungen erfüllt. Der Entscheidungsbaum empfiehlt somit, das k-Medoids Verfahren zur Behandlung von Ausreißern.

4.4 Evaluation und Diskussion der Ergebnisse

In diesem Abschnitt werden die Ergebnisse der vorangegangenen Kapitel in Bezug auf die Forschungsfrage zusammengefasst und diskutiert. Außerdem werden die vorhandenen Beschränkungen der Arbeit dargelegt. Die Vor- und Nachteile des entwickelten Entscheidungsbaumes werden ebenfalls dargestellt und diskutiert. Außerdem erfolgt eine kritische Auseinandersetzung mit der geleisteten Arbeit.

In dieser Arbeit wurden mit Hilfe einer strukturierten Literaturrecherche Verfahren ermittelt, welche im Zuge der Datenvorverarbeitung für das Data Mining Ausreißer behandeln können. Im Anschluss wurden die vorgestellten Verfahren in einem Entscheidungsbaum dargestellt, um den Entscheidungsprozess bezüglich einer Verfahrensauswahl unterstützen zu können. Mit Hilfe des in Kapitel 2 beschriebenen Stands der Technik konnten Schlüsselbegriffe in Bezug auf die Behandlung von Ausreißern abgeleitet werden, welche eine der Grundlagen der strukturierten Literaturrecherche gebildet haben. Durch die Suchbegriffe im Einzelnen und mehreren Kombinationen aus diesen konnten im Rahmen der durchgeführten strukturierten Literaturrecherche diverse Quellen ermittelt werden. Durch eine daraufhin durchgeführte sogenannte Vorwärts- und Rückwärtssuche wurden die in Tabell A-2 aufgelisteten Literaturwerke ermittelt. Aus diesen Quellen konnten Verfahren zur Behandlung von Ausreißern identifiziert werden. Diese sind die Clusteranalyse, Regressionsverfahren, Klassifikationsverfahren und Trimmen sowie das Winsorisieren. Werden die Ergebnisse der strukturierten Literaturrecherche mit dem Stand der Technik verglichen, können diese Verfahren größtenteils in Abschnitt 2.5 wiedergefunden werden. Ausschließlich das Trimmen wird im Stand der Technik dieser Arbeit nicht benannt. In der strukturierten Literaturrecherche konnten ebenfalls im Verhältnis zu den anderen Verfahren weniger Quellen zu dem Verfahren des Trimmens gefunden werden. Es ist anzumerken, dass noch weitere Möglichkeiten zur Behandlung von Ausreißern existieren. Aufgrund einer geringen Informationsdichte zu der Eignung zur Behandlung von Ausreißern wurden diese im Rahmen dieser Arbeit nicht berücksichtigt. Aus den in Abschnitt

2.3 besprochenen Phasen der Wissensentdeckung geht weiterhin hervor, dass die gefundenen Verfahren in unüberwachte und überwachte Verfahren eingeteilt werden können. Dies ergibt die strukturierte Literaturrecherche ebenfalls. Es kann außerdem festgestellt werden, dass die Einteilung in unüberwachte und überwachte Verfahren keine Rückschlüsse darauf ziehen lässt, auf welche Weise die Ausreißer behandelt werden. Sowohl die Clusteranalyse, welche zu den unüberwachten Verfahren zählt, als auch die Klassifikationsverfahren, welche zu den überwachten Verfahren zählen, identifizieren zunächst die Ausreißer, damit sie im Anschluss aus dem Datensatz entfernt werden können. Es konnte keine Literatur gefunden werden, die eine Übersicht über alle Verfahren gibt. Sharafi (2012) zum Beispiel bespricht in seinem Literaturwerk die Klassifikation und das Clustern. Die Verfahren des Trimmens und des Winsorisierens sowie die Regression werden hier nicht behandelt. Andere Autoren, wie zum Bespiel Chambers et al. (2000), behandeln nur ein einzelnes Verfahren. Bei diesen Quellen handelt es sich häufig um Zeitschriftenaufsätze. Außerdem kann durch die Betrachtung der in Tabelle A-2 aufgelisteten Literaturwerke erkannt werden, dass die meisten Quellen die Verfahren der Clusteranalyse und der Regression beschreiben. Beide Verfahren werden in Bezug auf die Behandlung von Ausreißern häufig vorgeschlagen. Aus der Datenvorverarbeitung in Abschnitt 2.4 geht hervor, dass nach Runkler (2020) Fehler in zufällige und systematische Fehler unterteilt werden können. Rauschen gehört nach Runkler (2020) zu den zufälligen Fehlern. Ausreißer können zusätzlich zu den zufälligen Fehlern auch zu den systematischen Fehlern gezählt werden. In der strukturierten Literaturrecherche in Kapitel 3 konnte festgestellt werden, dass viele Quellen die Begriffe des Ausreißers und des Rauschens nicht trennen. Daraus lässt sich ableiten, dass zur Behandlung von Ausreißern auch Verfahren betrachtet werden sollten, welche sich zunächst auf die Rauschbehandlung beziehen. Im Zuge dessen sollte immer untersucht werden, ob der jeweilige Autor des Literaturwerkes die Begriffe der Ausreißer und des Rauschen trennt oder nicht. Im Allgemeinen kann festgehalten werden, dass es die teilweise nicht vorhandene Trennung der Begriffe erschwert zu erkennen, ob ein Verfahren für die Behandlung von Ausreißern, Rauschen oder beidem geeignet ist. Die in Abschnitt 2.5 beschriebenen Punkt-, kontextualen und kollektiven Ausreißer werden in der anschließenden Literaturrecherche von dem größten Teil der Autoren nicht aufgegriffen. Die Kategorisierung nach der Art des Ausreißers ist allerdings sinnvoll, um die Behandlungsverfahren einordnen zu können. Daher wird angeregt, in zukünftigen Werken bezüglich der Ausreißerbehandlung in der Datenvorverarbeitung die Art des Ausreißers stärker mit einzubeziehen. Dadurch kann genauer eingegrenzt werden, für welche Datensätze die verschiedenen Verfahren aus Kapitel 3 geeignet sind. Des Weiteren ist aufgefallen, dass viele Literaturwerke in ihren Beschreibungen ausführlich auf die Identifizierung von Ausreißern eingehen. Die anschließende Behandlung wird allerdings in vielen Werken nur kurz erläutert. Die aus Kapitel 2 abgeleiteten Suchbegriffe in Tabelle 2 beinhalten zwar nicht die Identifizierung, allerdings die Begriffe Methode und Technik. Diese können neben der Behandlung auch mit der Identifizierung in Zusammenhang gebracht werden. In zukünftigen strukturiert durchgeführten Literaturrecherchen sollte daher darauf geachtet werden, dass genauer definierte Schlüsselbegriffe verwendet werden. In der strukturierten Literaturrecherche konnten einige Verfahren zur Behandlung von Ausreißern in der Datenvorverarbeitung ermittelt werden, welche als Data Mining Verfahren gelten. Daraus kann abgeleitet werden, dass sich Data Mining Verfahren, welche in dieser Arbeit nach dem CRISP-DM Modell im Anschluss an die Datenvorverarbeitung eingeordnet sind, ebenfalls zur Ausreißerbehandlung eignen. In Abschnitt 3.1 wurde deutlich, dass sich die Autoren der Quellen in Bezug auf die Eignung des k-Means Verfahrens für die Behandlung von Ausreißern teilweise uneinig sind. Olson und Lauhoff (2023), welche als Einzige das Verfahren als grundsätzlich geeignet einordnen, merken ebenfalls an, dass eine noch bessere Eignung erreicht werden kann, wenn der Mittelwert mit dem Median ersetzt wird. Obwohl das k-Means Verfahren zur Ausreißerbehandlung von einem Großteil der Autoren im Allgemeinen nicht empfohlen wurde, wird es in den Entscheidungsbaum mit einbezogen. Das ist darin begründet, dass es weitere fallspezifische Anforderungen und Kriterien geben kann. Durch diese kann es möglich sein, dass das k-Means Verfahren für den speziellen Anwen-

dungsfall geeignet ist. Insgesamt konnten in Kapitel 3 wenig Verfahren ermittelt werden, welche sich für die Behandlung von Ausreißern in nominal skalierten Daten eignen. Da diese in der Praxis jedoch ebenfalls relevant sind, sollten diese in zukünftigen Arbeiten detaillierter betrachtet werden. Des Weiteren konnten durch die strukturierte Literaturrecherche einige Einschränkungen der Verfahren ermittelt werden. Allerdings ist die Anzahl der gefundenen Einschränkungen in den Literaturwerken als relativ gering einzuordnen. Für die Entwicklung der Entscheidungsunterstützung in Abschnitt 4.3 ist es wichtig, genügend Informationen über die Beschränkungen eines Verfahrens zu haben um dieses fachlich korrekt in den Entscheidungsbaum einordnen zu können. Es ist festzuhalten, dass bei einer höheren Informationsdichte mehr Anforderungen und Kriterien für den entwickelten Entscheidungsbaum hätten abgeleitet werden können. Mit einer höheren Anzahl Kriterien könnte der Entscheidungsbaum detaillierter erstellt werden und somit eine bessere Unterstützung in der Entscheidungsfindung darstellen.

An den Abbildungen 15 und 16 und mit dem Fallbeispiel in Abschnitt 4.3 lassen sich die Vor- und Nachteile der Methode zur Entscheidungsunterstützung erkennen. Als großer Vorteil ist die Übersichtlichkeit des in dieser Arbeit entwickelten Entscheidungsbaumes zu sehen. Die Kriterien sind schnell zu erfassen und auch die Behandlungsverfahren lassen sich gut erkennen. Dadurch lassen sich die grundlegenden Eigenschaften übersichtlich darstellen. Dazu trägt auch die einfache Struktur bei, die ein Entscheidungsbaum mit sich bringt. Die Interpretation des in dieser Arbeit entwickelten Entscheidungsbaumes ist leicht und verständlich. Wie in Abschnitt 2.1 beschrieben wurde, soll eine Methode zur Entscheidungsunterstützung ebenfalls dabei unterstützen, das Problem zu strukturieren. Das ist mit einem Entscheidungsbaum durch die Einteilung in Kriterien gut umzusetzen. Außerdem lässt sich der Entscheidungsbaum auch auf andere Problem- bzw. Fragestellungen anpassen. Der entwickelte Entscheidungsbaum ist darüber hinaus auch geeignet für Erweiterungen. Bereits bestehende, aber in dieser Arbeit nicht betrachtete, Verfahren oder neu entwickelte Verfahren zur Behandlung von Ausreißern können zum jeweiligen Blatt hinzugefügt werden. Außerdem können weitere Kriterien oder Anforderungen zu denen aus Abschnitt 4.1 und 4.2 hinzugefügt werden, indem weitere Zweige und Knoten zu den bereits bestehenden hinzugefügt werden. Ein weiterer Vorteil des Entscheidungsbaumes ist, dass er sich schrittweise aufbauen lässt. Dadurch ergibt sich eine klare Struktur, an welcher sich der Verwender des Entscheidungsbaums orientieren kann. Mit Hilfe der graphischen Darstellung können die komplexen Entscheidungen in Bezug auf die Verfahrensauswahl anschaulich dargestellt werden. Die gewählte Methode zur Entscheidungsunterstützung erleichtert es außerdem dem Anwender, seine Entscheidung zur Behandlung von Ausreißern nachvollziehbar zu erläutern. Darüber hinaus ist es mit einem geringen Aufwand möglich, den entwickelten Entscheidungsbaum zu digitalisieren.

Gegenüber den genannten Vorteilen hat der entwickelte Entscheidungsbaum auch einige Nachteile. Bei einigen Kombinationen aus Eigenschaften des zu untersuchenden Datenbestandes schlägt der Entscheidungsbaum mehr als ein Verfahren zu Behandlung der Ausreißer vor. In diesen Fällen muss von dem Anwender des Verfahrens eine abschließende Entscheidung getroffen werden. Dies setzt voraus, dass der Anwender eine gewisse Erfahrung mit sich bringt und entscheiden kann, welches der vorgeschlagenen Verfahren er abschließend auswählt. Diese Problematik könnte abgeschwächt werden, indem praktische Versuche durchgeführt werden. Aus diesen könnten Erkenntnisse erlangt und neue Anforderungen an die Verfahren abgeleitet werden, die im Rahmen des Entscheidungsbaumes eine detailliertere Unterscheidung zwischen den Verfahren ermöglichen. Außerdem ist es problematisch, dass keine genaue Angabe darüber getroffen werden kann, ab wann eine Datenmenge als groß gilt. Um bei diesem Kriterium eine bessere Zuordnung ermöglichen zu können, können ebenfalls praktische Versuche durchgeführt werden, mit deren Hilfe eine eindeutige Einordnung vorgenommen werden kann. Dies stellt eine inhaltliche Lücke in den gefundenen Quellen der strukturierten Literaturrecherche dar. Des Weiteren beruht die entwickelte Methode zur Entscheidungsunterstützung auf vier Kriterien. Diese vier Kriterien haben sich aus Kapitel 3 ergeben.

Es gibt jedoch weitere Kriterien, die berücksichtigt werden können. Um in der Praxis eine fundierte Entscheidung treffen zu können, kann der in dieser Arbeit entwickelte Entscheidungsbaum keine allumfassende Entscheidungsempfehlung geben. Zum Beispiel ist es in diesem Entscheidungsbaum nicht vorgesehen, dass mehr als ein Skalenniveau in einem Datenbestand vorkommt. Wie in der strukturierten Literaturrecherche in Kapitel 3 aber festgestellt wurde, kommen in der Praxis auch Datenbestände vor, welche nominale und numerische Attribute enthalten. Diese Eigenschaft eines Datenbestandes wird in dem entwickelten Entscheidungsbaum nicht abgebildet, da nicht genügend Informationen über alle Verfahren vorliegen. Es wurden in Kapitel 4 nur Kriterien berücksichtigt, bei denen die Informationsdichte über alle Verfahren genügend groß war. Es existieren allerdings wie bereits erwähnt weitere Kriterien, die fallspezifische Auswirkungen auf die Auswahl eines Behandlungsverfahren für Ausreißer haben. Daneben ist es durch die Wahl der Methode eines Entscheidungsbaums nicht möglich, die Anforderungen an die verschiedenen Verfahren zu gewichten. Dies kann aber unter Umständen für eine Anwendung in der Praxis ausschlaggebend sein. Es können Anforderungen existieren, die für den speziellen Fall des vorliegenden Datenbestandes wichtiger sind als andere. Daher könnte es sinnvoll für den Anwender sein, eine Entscheidungsunterstützung nutzen zu können, welche das Gewichten von Kriterien ermöglicht. Um mit Hilfe des Entscheidungsbaumes detailliertere Abgrenzungen der Verfahren festzulegen, werden weitere Kriterien benötigt. Die Integration dieser in den entwickelten Entscheidungsbaum würde wie oben bereits beschrieben zur Folge haben, dass die Methode zur Entscheidungsunterstützung nach einer gewissen Anzahl an weiteren Kriterien unübersichtlich werden kann.

In Bezug auf die durchgeführte Literaturrecherche in Kapitel 3 sowie die Methode zur Entwicklung einer Entscheidungsunterstützung in Kapitel 4 wird deutlich, dass diese Arbeit einige Beschränkungen hat. In dieser Arbeit wurden nur die Regressions-, Cluster- und Klassifikationsverfahren sowie Trimmen und Winsorisieren betrachtet. Es konnten weitere Verfahren durch die strukturierte Literaturrecherche ermittelt werden, wie zum Beispiel die Assoziationsanalyse, welche keinem der zuvor genannten zuzuordnen sind. Allerdings war die Informationsdichte bezüglich anderer Verfahren nicht ausreichend hoch. Außerdem ist es für die Entwicklung der in Abschnitt 4.3 beschriebenen Methode zur Entscheidungsunterstützung erforderlich, neben einer umfangreichen Literaturrecherche auch praktische Versuche durchzuführen, um die Gemeinsamkeiten und Unterschiede der beschriebenen Verfahren umfassend herauszuarbeiten und somit in den Entscheidungsbaum mit einzubeziehen. Auf Grund des Umfangs dieser Arbeit wurden keine praktischen Versuche durchgeführt. Trotzdem können aus dieser Arbeit einige wichtige Erkenntnisse abgeleitet werden. Es wurde gezeigt, dass das Winsorisieren und Trimmen geeignete Verfahren zur Behandlung von Ausreißern darstellen, auch wenn sie nicht weit verbreitet sind und selten genutzt werden. Darüber hinaus wurde deutlich, dass die Verfahren auch danach eingeteilt werden können, auf welche Art und Weise sie mit Ausreißern umgehen. In dieser Arbeit wurden die Möglichkeiten der Entfernung von Ausreißern, das Ersetzen von Ausreißern sowie den Einfluss der Ausreißer verringern berücksichtigt. Die entwickelte Methode zur Entscheidungsunterstützung in Form eines Entscheidungsbaums ist eine gute Grundlage, um auf ihr aufzubauen und sie weiter zu entwickeln. Sie stellt einen geeigneten Ausgangspunkt dar, um Verfahren einordnen zu können und eine gute Übersicht zu schaffen. Zusammenfassend ist zu sagen, dass in dieser Arbeit größtenteils die häufig vorkommenden Verfahren zur Behandlung von Ausreißern ausgemacht und ihren Grundlagen erläutert wurden.

Das Ergebnis dieser Arbeit ist, dass der entwickelte Entscheidungsbaum grundsätzlich dazu in der Lage ist, den Prozess der Ausreißerbehandlung in der Datenvorverarbeitung für das Data Mining zu unterstützen. Die Übersichtlichkeit der ausgewählten Methode zur Entscheidungsunterstützung und die einfach Nachvollziehbarkeit sind hervorzuheben. Um den entwickelten Entscheidungsbaum auch in der Praxis anwenden zu können, müsste zunächst durch praktische Versuche definiert werden, wann eine Datenmenge als klein und wann als groß gilt. Außerdem sollten Kombinationen der Ausprägungen der bisher verwendeten Kriterien möglich sein, wie zum Beispiel Datenbestände mit ordinal- und metrischskalierten Daten. Zusätzlich

sollten weitere Kriterien mit in den Entscheidungsbaum einbezogen werden. Diese Kritikpunkte müssen behandelt werden, um die Verwendung des Entscheidungsbaums in der Praxis zu ermöglichen. Die entwickelte Methode bietet eine gute Grundlage, um auf ihr aufbauend den Entscheidungsbaum zu erweitern. In dem jetzigen Zustand des Entscheidungsbaumes ist er nicht geeignet, um in der Praxis angewandt zu werden.

5 Zusammenfassung und Ausblick

Das Hauptziel dieser Arbeit war es, eine strukturierte Analyse von Verfahren zur Behandlung von Ausreißern in der Datenvorverarbeitung für das Data Mining durchzuführen und anschließend eine Methode zur Entscheidungsunterstützung zu entwickeln. Dazu wurde in Kapitel 2 der Stand der Technik erläutert. Dieser bestand unter anderem aus den Phasen der Wissensentdeckung und einer allgemeinen Einführung in die Datenvorverarbeitung. Außerdem wurden in diesem Kapitel die grundsätzlichen Begriffe und Definitionen in Bezug auf Ausreißer thematisiert, verschiedene Arten von Ausreißern vorgestellt und kurz auf die Behandlungsmöglichkeiten von Ausreißern eingegangen. Im anschließenden Kapitel 3 wurde zunächst die Vorgehensweise der strukturierten Literaturrecherche beschrieben. Dazu wurde das Vorgehen nach vom Brocke et al. (2009) vorgestellt und angewandt. Die Schlüsselbegriffe konnten aus Kapitel 2 abgeleitet werden. Außerdem wurde in diesem Abschnitt der Rechercheumfang zur Durchführung der systematischen Literaturrecherche festgelegt. Das Ergebnis der strukturierten Literaturrecherche wurde in einer Übersicht über die gefundene Literatur in Form einer Tabelle dargestellt. Die gesamte Tabelle A-2 wurde im Anhang abgebildet. Im Anschluss daran wurden die gefundenen Verfahren nach übergeordneten Kategorien sortiert und vorgestellt. Dazu wurde jeweils ein Auszug aus Tabelle A-2 abgebildet. Darauf folgend wurden verschiedene Verfahren der jeweiligen Kategorie erläutert. Außerdem wurden zum Ende jedes Abschnitts die Ansichten bezüglich der Eignung zur Ausreißerbehandlung der Autoren gegenübergestellt. Auf der Grundlage der gewonnenen Erkenntnisse aus Kapitel 3 und mit Hilfe von Kapitel 2 wurde im darauffolgenden Kapitel 4 eine Methode zur Entscheidungsunterstützung entwickelt. Dafür wurden in Abschnitt 4.1 zunächst Anforderungen an die vorgestellten Verfahren abgeleitet. Aus diesen konnten in Abschnitt 4.2 Kriterien formuliert werden, welche die Grundlage für die Methode zur Entscheidungsunterstützung darstellten. Die Kriterien waren die Datenmenge, das Skalenniveau, die Art des Ausreißers sowie die Art des Datensatzes. Daraus konnte anschließend in Abschnitt 4.3 der Entscheidungsbaum entwickelt werden. Nach Festlegen einer geeigneten Abfolge der Kriterien im Entscheidungsbaum wurden die gefundenen Verfahren zu den jeweiligen Ausprägungen der Kriterien zugeordnet. Die Kriterien konnten entlang der Zweige die Entscheidung zur Verfahrenswahl zur Behandlung von Ausreißern unterstützt werden. Dabei konnte festgehalten werden, dass bei bestimmten Voraussetzungen mehrere Verfahren als Behandlungsmöglichkeit vom Entscheidungsbaum vorgeschlagen wurden.

Es ist zu erkennen, dass die entwickelte Methode zur Entscheidungsunterstützung eine Grundlage zur Verfahrensauswahl darstellen und den Entscheidungsprozess unterstützen kann, wenn sie verfeinert und erweitert wird. Für eine praktische Anwendung sollten weitere Anforderungen und Kriterien formuliert werden, um eine bessere und eindeutigere Entscheidung in Bezug auf die Ausreißerbehandlung treffen zu können. Zusätzlich sollten praktische Versuche durchgeführt werden. Dies wäre unter Anderem sinnvoll um festlegen zu können, ab wann eine Datenmenge als groß gilt. Dies konnte im Zuge der strukturierten Literaturrecherche nicht ermittelt werden. Ein weiterer Vorteil der Durchführung von Praxistests wäre, detaillierteres Wissen über die Verfahren zu erlangen. Da die strukturierte Literaturrecherche gezeigt hat, dass auch Verfahren zur Behandlung von Rauschen für die Behandlung von Ausreißern verwendet werden können, sollte diesem Bereich ebenfalls weiter nachgegangen werden.

Literaturverzeichnis

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

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

Aggarwal, Charu C.; Sathe, Saket (2017): Outlier Ensembles. An Introduction. 1st ed. 2017. Cham: Springer International Publishing; Imprint: Springer.

Aguinis, Herman; Gottfredson, Ryan K.; Joo, Harry (2013): Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. In: *Organizational Research Methods* 16 (2), S. 270–301. DOI: 10.1177/1094428112470848.

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

Alpaydin, Ethem (2019): Maschinelles Lernen. 2. Auflage. Berlin, Boston: De Gruyter Oldenbourg (De Gruyter Studium).

Andreopoulos, Bill; An, Aijun; Wang, Xiaogang; Schroeder, Michael (2009): A Roadmap of Clustering Algorithms: Finding a Match for a Biomedical Application. In: *Briefings in bioinformatics* 10 (3), S. 297–314. DOI: 10.1093/bib/bbn058.

Bacher, Johann (2010): Clusteranalyse. Anwendungsorientierte Einführung in Klassifikationsverfahren. 3., erg., vollst. überarb. u. neu gestalt. Aufl. München: Oldenbourg. Online verfügbar unter https://www.degruyter.com/isbn/9783486710236.

Bell, David E.; Raiffa, Howard; Tversky, Amos (1988): Descriptive, Normative and Prescriptive Interactions in Decision Making 1988, S. 9–30. DOI: 10.1017/CBO9780511598951.003.

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

Blaine, Bruce E. (2018): Winsorizing. In: *The Sage encyclopedia of educational research, measurement, and evaluation*, S. 1817–1818. DOI: 10.4135/9781506326139.

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

Bramer, Max (2016): Principles of Data Mining. 3rd ed. 2016. London: Springer London; Imprint: Springer (Undergraduate Topics in Computer Science).

Breunig, Markus M; Kriegel, Hans-Peter; Ng, Raymond T.; Sander, Jörg (2000): LOF: Identifying Density-Based Local Outliers. In: *ACM Sigmod Record*, Artikel 93-104.

Brinkmeyer, Dieter; Müller, Rolf A. E. (1994): Entscheidungsunterstützung mit dem AHP. In: *Zeitschrift für Agrarinformatik* (5), S. 82–92.

Buchholz, Peter; Clausen, Uwe (2009): Grosse Netze der Logistik. Die Ergebnisse des Sonderforschungsbereichs 559. 1. Aufl. Berlin: Springer.

Chandola, Varun; Banerjee, Arindam; Kumar, Vipin (2009): Anomaly Detection. In: *ACM Comput. Surv.* 41 (3), S. 1–58. DOI: 10.1145/1541880.1541882.

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

Cohen, Jacob; Aiken, Leona S.; Cohen, Patricia; West, Stephen G. (2003): Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. 3. Aufl.: Lawrence Erlbaum Associates.

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

Divya, D; Babu, Suvanam Sasidhar (2016): Methods to Detect Different Types of Outliers. In: *International Conference on Data Mining and Advanced Computing*, S. 23–28. DOI: 10.1109/SAPIENCE.2016.7684114.

D'Onofrio, Sara; Meier, Andreas (Hg.) (2021): Big Data Analytics. Grundlagen, Fallbeispiele und Nutzungspotenziale. Wiesbaden: Springer Vieweg (Springer eBook Collection).

Ester, Martin; Sander, Jörg (2000): Knowledge Discovery in Databases. Techniken und Anwendungen. Berlin, Heidelberg: Springer.

Fasel, Daniel; Meier, Andreas (Hg.) (2016): Big data. Grundlagen, Systeme und Nutzungspotenziale. Springer Fachmedien Wiesbaden. Wiesbaden: Springer Vieweg (Praxis der Wirtschaftsinformatik).

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

Gama, Joao (2010): Knowledge Discovery from Data Streams. 1. Aufl. USA: Chapman & Hall/CRC, Boca Raton.

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

Ghavami, Peter (2020): Big Data Analytics Methods. Analytics Techniques in Data Mining, Deep Learning and Natural Language Processing. 2nd edition. Boston, Berlin: De Gruyter (Business & economics). Online verfügbar unter https://www.degruyter.com/isbn/9781547401567.

Giloni, Avi; Simonoff, Jeffrey S.; Sengupta, Bhaskar (2006): Robust Weighted LAD Regression. In: *Computational Statistics & Data Analysis* 50 (11), S. 3124–3140. DOI: 10.1016/j.csda.2005.06.005.

Gosh, Dhiren; Vogt, Andrew (2012): Outliers: An Evaluation of Methodologies. In: *Joint Statistical Meetings*, S. 3455–3460. Online verfügbar unter http://www.asasrms.org/Proceedings/y2012/Files/304068_72402.pdf.

Gupta, Manish; Gao, Jing; Aggarwal, Charu; Han, Jiawei (2014): Outlier Detection for Temporal Data. 1st ed. 2014. Cham: Springer International Publishing; Imprint Springer (Synthesis Lectures on Data Mining and Knowledge Discovery).

Han, Jiawei; Kamber, Micheline; Pei, Jian (2012): Data Mining. Concepts and Techniques. 3. Aufl. Waltham: Morgan Kaufmann.

Hawkins, D. M. (1980): Identification of Outliers. Dordrecht: Springer (Springer eBook Collection Mathematics and Statistics).

Hoo, K. A.; Tvarlapati, K. J.; Pivoso, M. J.; Hajare, R. (2002): A Method of Robust Multivariate Outlier Replacement. In: *Computers and Chemical Engineering* (26), S. 17–19.

Hotho, Andreas (2004): Clustern mit Hintergrundwissen. Doktorarbeit. Universität Fridericiana zu Karlsruhe.

Huber, Peter J.; Ronchetti, Elvezio M (2009): Robust Statistics. 2. Aufl. New Jersey: John Wiley & Sons Incorporated.

ISO/IEC 2382:2015. Online verfügbar unter https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v2:en, zuletzt geprüft am 02.04.2023.

Joenssen, Dieter William; Müllerleile, Thomas (2014): Fehlende Daten beim Data-Mining. In: *HMD* 51 (4), S. 458–468. DOI: 10.1365/s40702-014-0038-8.

Kik, David (2022): Zur modellbasierten Entscheidungsunterstützung von Unternehmen in der regionalen Standortplanung und -entwicklung. Wiesbaden, Heidelberg, Wiesbaden, Heidelberg: Springer Gabler; Imprint Springer Gabler (Produktion und Logistik). Online verfügbar unter http://www.springer.com/.

Koufakou, Anna; Georgiopoulos, Michael (2010): A Fast Outlier Detection Strategy for Distributed High-dimensional Data Sets with Mixed Attributes. In: *Data Min Knowl Disc* 20 (2), S. 259–289. DOI: 10.1007/s10618-009-0148-z.

Kureljusic, Marko; Karger, Erik (2022): Data Preprocessing as a Service – Outsourcing der Datenvorverarbeitung für KI-Modelle mithilfe einer digitalen Plattform. In: *Informatik Spektrum* 45 (1), S. 13–19. DOI: 10.1007/s00287-021-01420-5.

Lee, Jae Hyuk; Park, Jeong Jun; Yoon, Hyungchul (2020): Automatic Bridge Design Parameter Extraction for Scan-to-BIM. In: *Applied Sciences* 10 (20), S. 7346. DOI: 10.3390/app10207346.

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

Meier, Andreas; Kaufmann, Michael (2016): SQL- & NoSQL-Datenbanken. 8., überarbeitete und erweiterte Auflage. Berlin, Heidelberg: Springer Vieweg (eXamen.press). Online verfügbar unter http://www.springer.com/.

Mishra, Puneet; Biancolillo, Alessandra; Roger, Jean Michel; Marini, Federico; Rutledge, Douglas N. (2020): New Data Preprocessing Trends Based on Ensemble of Multiple Preprocessing Techniques. In: *TrAC Trends in Analytical Chemistry* 132, S. 116045. DOI: 10.1016/j.trac.2020.116045.

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

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

Olson, David L.; Lauhoff, Georg (2023): Deskriptives Data-Mining. Cham: Springer Nature Switzerland; Imprint Springer.

Petersohn, Helge (2005): Data Mining. Verfahren, Prozesse, Anwendungsarchitektur. Zugl.: Leipzig, Univ., Habil.-Schr., 2004. München, Wien: Oldenbourg. Online verfügbar unter http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&doc_number=013355129&line_number=0002&func_code=DB_RECORDS&service_type=MEDIA.

Radovanovic, Milos; Nanopoulos, Alexandros; Ivanovic, Mirjana (2015): Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection. In: *IEEE Trans. Knowl. Data Eng.* 27 (5), S. 1369–1382. DOI: 10.1109/TKDE.2014.2365790.

Rajeswari, A. M.; Yalini, S. K.; Janani, R.; Rajeswari, N.; Deisy, C. (2018): A Comparative Evaluation of Supervised and Unsupervised Methods for Detecting Outliers. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). Coimbatore, India, 4/20/2018 - 4/21/2018: IEEE, S. 1068–1073.

Ranga Suri, N. N. R.; Athithan, G.; Murty M, Narasimha (2019): Outlier Detection Techniques and Applications. A Data Mining Perspective. 1st ed. 2019. Cham: Springer International Publishing; Imprint: Springer (Intelligent Systems Reference Library, 155).

Romão, Xavier; Vasanelli, Emilia (2021): Identification and Processing of Outliers. In: Denys Breysse und Jean-Paul Balayssac (Hg.): Non-Destructive In Situ Strength Assessment of Concrete, Bd. 32. Cham: Springer International Publishing (RILEM State-of-the-Art Reports), 161-180.

Runkler, Thomas A. (2000): Information Mining. Methoden, Algorithmen und Anwendungen intelligenter Datenanalyse. Wiesbaden: Vieweg+Teubner Verlag (Computational Intelligence).

Runkler, Thomas A. (2020): Data Analytics. Models and Algorithms for Intelligent Data Analysis. 3. Aufl.: Springer Fachmedien Wiesbaden.

S. Baumann; M. Gnisia; et al. (2018): Identifikation und Behandlung von Ausreißern in Flugbetriebsdaten für Machine Learning Modelle.

Sharafi, Armin (2012): Knowledge Discovery in Databases. Dissertation. Technische Universität München, München.

Spengler, Thomas; Rommelfanger, Heinrich; Geiger, Martin Josef; Metzger, Olga; Fichtner, Wolf (Hg.) (2017): Entscheidungsunterstützung in Theorie und Praxis. Tagungsband zum Workshop FEU 2016 der Gesellschaft für Operations Research e.V. Wiesbaden: Springer Fachmedien Wiesbaden. Online verfügbar unter https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=4835450.

Vom Brocke, Jan; Simons, Alexander; Niehaves, Bjoern; Reimer, Kai; Plattfaut, Ralf; Cleven, Anne (2009): Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process. In: *In Proceedings of the 17th European Conference on Information Systems*, S. 2206–2217.

Wirth, Rüdiger; Hipp, Jochen (2000): CRISP-DM: Towards a Standard Process Model for Data Mining. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery in databases* (1), S. 29–39. Online verfügbar unter http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.

Zhang, Haofan; Nian, Ke; Coleman, Thomas F.; Li, Yuying (2020): Spectral Ranking and Unsupervised Feature Selection for Point, Collective, and Contextual Anomaly Detection. In: *Int J Data Sci Anal* 9 (1), S. 57–75. DOI: 10.1007/s41060-018-0161-7.

Anhang A

Tabelle A-1 Beispielhafter Suchvorgang in der Datenbank Scopus

Suchfeld	Schlüsselbegriff	Anzahl der Ergebnisse	Operator
All fields	Wissensentdeckung	498	
All fields	Wissensentdeckung, Ausreißer	0	AND
All fields	data preprocessing, outlier	5527	AND
Article title, Abstract, Keywords	data preprocessing, outlier	1195	AND
Article title, Abstract, Keywords	data preprocessing, outlier, treatment	53	AND
Article title, Abstract, Keywords	data preprocessing, outlier, handling	199	AND

Fortsetzung Tabelle A-1

Article title, Abstract, Keywords	outlier, techniques	13528	AND
Article title, Abstract, Keywords	outlier, techniques, handling	1007	AND
Article title, Abstract, Keywords	outlier, winsorization	52	AND
Article title, Abstract, Keywords	outlier, winsorization, handling	1	AND
Article title	outlier, handling	61	AND

Tabelle A-2 Übersicht der Ergebnisse der Ergebnisse der strukturierten Literaturrecherche

Autor (Jahr)	Titel	Behandlung von Ausreißern	Verfahren
Aggarwal (2015)	Data Mining	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren
Aggarwal (2017)	Outlier Analysis	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren
Aggarwal und Sathe (2017)	Outlier Ensembles	Entfernen, Einfluss verringern	Trimmen und Winsorisieren
Agostinelli et al. (2015)	Robust Estimation of Multivariate Location and Scatter in the Presence of Cellwise and Casewise Contamination	Ersetzen	Regressionsverfahren
Aguinis et al. (2013)	Best-Practice Recommendations for Defining, Identifying and Handling Outliers	Ersetzen, Entfernen, Einfluss Verringern	Clusterverfahren, Regressionsverfahren, Trimmen und Winsorisieren
Alasadi und Bhaya (2017)	Review of Data Preprocessing Techniques	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren

Fortsetzung Tabelle A-2

Alpaydin (2019)	Maschinelles Lernen	Ersetzen, Entfernen	Regressionsverfahren, Klassifikation
Andreopoulos et al. (2009)	A Roadmap of Clustering Algorithms	Entfernen	Clusterverfahren
Bacher (2010)	Clusteranalyse	Entfernen	Clusterverfahren
Barbará (2002)	Requirements for Clustering Data Streams	Entfernen	Clusterverfahren
Barnett (2004)	Environmental Statistics	Ersetzen, Entfernen, Einfluss verringern	Regressionsverfahren, Trimmen und Winsorisieren
Blaine (2018)	Winsorizing	Entfernen, Einfluss verringern	Trimmen und Winsorisieren
Bramer (2016)	Principles of Data Mining	Entfernen	Klassifikation

Fortsetzung Tabelle A-2

Chambers et al. (2000)	Winsorization for Identifying and Treating Outliers in Business Surveys	Einfluss verringern	Winsorisieren
Chen und Wang (2008)	The Data Mining Technology Based on CIMS and its Application on Automotive Remanufacturing	Entfernen	Clusterverfahren
Cleve und Lämmel (2020)	Data Mining	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren, Klassifikation
Cohen et al. (2003)	Applied Multiple Regression/Correlation Analysis for the Behavioral Sience	Ersetzen	Regressionsverfahren
D'Onofrio und Meier (2021)	Big Data Analytics	Entfernen	Clusterverfahren
Ester und Sander (2000)	Knowledge Discovery in Databases	Entfernen	Clusterverfahren, Klassifikation
Gama (2010)	Knowledge Discovery from Datastreams	Ersetzen	Clusterverfahren

Fortsetzung Tabelle A-2

Ghavami (2020)	Big Data Analytics Methods	Ersetzen	Regressionsverfahren
Giloni et al. (2006)	Robust Weighted LAD Regression	Ersetzen	Regressionsverfahren
Gosh und Vogt (2012)	Outliers: An Evaluation of Methodologies	Entfernen, Einfluss verringern	Trimmen und Winsorisieren
Han et al. (2012)	Data Mining. Concepts and Techniques	Ersetzen	Regressionsverfahren
Hoo et al. (2002)	A Method of Robust Multivariate Outlier Replacement	Einfluss verringern	Winsorisieren
Hotho (2004)	Clustern mit Hintergrundwissen	Entfernen	Clusterverfahren
Huber und Ronchetti (2009)	Robust Statistics	Entfernen, Einfluss verringern	Trimmen und Winsorisieren

Fortsetzung Tabelle A-2

Jannaschk (2017)	Infrastruktur für ein Data Mining Design Framework	Entfernen	Clusterverfahren
Khan et al. (2007)	Robust Linear Model Selection Based on Least Angle Regression	Ersetzen, Entfernen, Einfluss verringern	Regressionsverfahren, Trimmen und Winsorisieren
Lee et al. (2020)	Automatic Bridge Design Parameter Extraction for Scan-to-BIM	Entfernen	Klassifikation
Luengo et al. (2020)	Big Data Preprocessing	Entfernen	Klassifikation
Mallikharjuna et al. (2023)	Data Preprocessing Techniques	Entfernen	Clusterverfahren
Olson und Lauhoff (2023)	Deskriptives Data-Mining	Entfernen	Clusterverfahren
Petersohn (2005)	Data Mining	Entfernen	Clusterverfahren

Fortsetzung Tabelle A-2

Radovanovic et al. (2015)	Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection	Entfernen	Klassifikation
Runkler (2020)	Data Analytics	Ersetzen, Entfernen	Clusterverfahren, Regressionsverfahren, Klassifikation
Sharafi (2012)	Knowledge Discovery in Databases	Entfernen	Clusterverfahren, Klassifikation
Sullivan et al. (2021)	So Many Ways To Assess Outliers	Entfernen, Einfluss verringern	Trimmen und Winsorisieren
Yi (2023)	Robust and Multivariate Statistical Methods	Entfernen, Einfluss verringern	Trimmen und Winsorisieren
Yuan und Bentler (1998)	Structural Equation Modeling with Robust Covariances	Ersetzen	Regressionsverfahren
Zhong und Yuan (2011)	Bias and Efficiency in Structural Equation Modeling	Ersetzen	Regressionsverfahren