Exemplarische Anwendung und Vergleich verschiedener Imputationsverfahren

als Vorbereitung für Data Mining in der Wissensentdeckung in Datenbanken

Masterarbeit zur Erlangung des Grades M. Sc.

Name:	Lukas Börger
Matrikelnummer:	195406
Studiengang:	Wirtschaftsingenieurwesen

Ausgabedatum:	18.11.2022
Abgabedatum:	30.05.2023

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

Technische Universität Dortmund

Fakultät Maschinenbau

Fachgebiet IT in Produktion und Logistik

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

KDD	Knowledge Discovery in Databases
KDID	Knowledge Discovery in Industrial Databases
MNAR	Missing not at random
MAR	Missing at random
MCAR	Missing completely at random
JM	Joint Modeling
FCS	Fully Conditional Specification
MVNI	Multivariate Normal Imputation
ML	Maximum Likelihood

1 Einleitung

Heutzutage werden in diversen Bereichen Daten gesammelt und mit zunehmender Geschwindigkeit angehäuft (Fayyad et al. 1996). Nachdem die Nutzung von Daten ursprünglich den empirischen Wissenschaften vorbehalten war, sind Daten mittlerweile auch für andere Branchen von großer Bedeutung. So können Daten im betriebswirtschaftlichen Sinne als Rohstoff verstanden werden, mithilfe dessen Wissen als entscheidender Wirtschaftsfaktor generiert werden kann (Plaue 2021). Doch während die Sammlung von Daten demnach eine Chance darstellen kann, so werden Unternehmen durch die ständig wachsende Datenmenge bei der zweckmäßigen Nutzung der Daten gleichzeitig auch vor Probleme gestellt (Fayyad 2005). Daher hat sich die Datenwissenschaft zur datengestützten Wissensgewinnung mittlerweile zu einer eigenständigen Forschungsdisziplin entwickelt (Smith 2006).

Zu ebenjener Wissensgenerierung aus Daten haben sich in den vergangenen Jahrzehnten verschiedene Vorgehensmodelle zur Wissensentdeckung in Datenbanken entwickelt, deren Kern in der Regel die Datenanalyse bzw. das Data Mining bildet (vgl. Brachman & Anand 1996; Fayyad et al. 1996). Neben der Bewältigung riesiger Datenmengen stellen Qualitätsmängel realer Datenbestände im Vorfeld der Datenanalyse eine der größten Herausforderungen dar. Dabei besteht die häufigste Ursache in fehlenden Merkmalswerten. In der Literatur wird davon ausgegangen, dass unvollständige Datensätze nicht die Ausnahme, sondern die Regel sind (Rockel 2017).

> *„Missing data are not problematic, per se – how we approach and treat missing data, on the other hand, can be highly problematic. “*
> (Little et al. 2014, S.151)

Das häufige Auftreten fehlender Daten und das hier vorangegangene Zitat verdeutlichen, dass dem Umgang mit fehlenden Daten im Kontext der Wissensentdeckung in Datenbanken eine große Bedeutung zugerechnet wird. Daher haben sich in der Vergangenheit parallel zur Entwicklung neuer Datenanalyseverfahren auch immer mehr Verfahren zum Umgang mit fehlenden Daten herausgebildet.

Diese Arbeit beschäftigt sich im Bereich des Umgangs mit fehlenden Werten speziell mit der Imputation von Werten. Imputationsverfahren umfassen Techniken zum Füllen fehlender Werte und stellen die gängigste Lösung für den Umgang mit fehlenden Merkmalswerten dar. Dementsprechend existieren mittlerweile diverse Veröffentlichungen, die Imputationsverfahren explizit thematisieren oder die Nutzung von Imputationsverfahren als Werkzeug aufgreifen. Dabei existieren auf der einen Seite diverse Veröffentlichungen, die Imputationsverfahren zwar

als Methode im Wissensentdeckungsprozess in Datenbanken benennen, diese darüber hinaus aber nicht näher behandeln. Auf der anderen Seite setzen Quellen, die explizit die Imputation fehlender Daten fokussieren, diese nur selten in den Gesamtkontext von Wissensentdeckungsprozessen. Weitergehend werden Imputationsverfahren entweder in Grundlagenwerken auf theoretischer Ebene vorgestellt oder als Werkzeug in Studien verwendet. Dabei wird die eigentliche Anwendung der Verfahren in Studien zumeist nur kurz oder gar nicht näher ausgeführt.

Diese Arbeit verfolgt daher das übergeordnete Ziel, Imputationsverfahren und deren Anwendung im Gesamtkontext der Wissensentdeckung in Datenbanken vergleichen zu können. Um dieser allgemeinen Zielsetzung gerecht zu werden, werden verschiedene, untergeordnete Ziele definiert. Zunächst wird das grundlegende Ziel verfolgt, Imputationsverfahren in den Gesamtkontext der Datenwissenschaften und der Wissensentdeckung in Datenbanken einzuordnen. Weitergehend wird dann die Vergleichbarkeit von Imputationsverfahren fokussiert. Dazu wird danach gefragt, wie sich Imputationsverfahren bewerten und miteinander vergleichen lassen. Sofern möglich, wird in diesem Zuge auch die Frage beantwortet, inwiefern verschiedene Imputationsverfahren und deren Ergebnisse die Leistungsfähigkeit anschließender Data-Mining-Prozesse und damit auch ganze Wissensentdeckungsprozesses beeinflussen.

Zur Verfolgung dieser Ziele und zur Beantwortung dieser Forschungsfragen führt diese Arbeit im Grundlagenteil zunächst in die Domäne der Datenwissenschaft und in die Thematik der Wissensentdeckung in Datenbanken ein. Im Bereich der Wissensentdeckung in Datenbanken werden dazu verschiedene Vorgehensmodelle präsentiert, um Imputationsverfahren gleichzeitig in den Ablauf einzuordnen und deren Stellenwert zu verdeutlichen. Dabei wird auch ein Bezug zu vor- und nachgelagerten Schritten, insbesondere dem Data Mining, hergestellt. Anschließend werden dann zunehmend Imputationsverfahren selbst fokussiert. Dazu werden in Aussicht auf den Hauptteil zunächst verschiedene Einteilungsmöglichkeiten von Imputationsverfahren und anschließend ausgewählte Imputationsverfahren bzw. Verfahrensgruppen selbst vorgestellt.

Im Hauptteil der Arbeit wird dann die Fragestellung nach der Vergleichbarkeit von Imputationsverfahren adressiert. Dazu wird im dritten Kapitel anhand einer systematische Literaturrecherche erörtert, wie sich Imputationsverfahren bewerten lassen. Dabei werden zunächst die Vorgehensweisen zum Vergleich der Verfahren analysiert, indem die Studien auf bestimmte Rahmenbedingungen, Einflussfaktoren, untersuchte Faktoren und Auswertungsmetriken untersucht werden. Als Ergebnis dieser Analyse wird schließlich ein Rahmenkonzept zum Vergleich von Imputationsverfahren vorgeschlagen, das die zuvor untersuchten Aspekte übersichtlich zusammenfasst und verknüpft.

Neben der Frage, wie Imputationsverfahren verglichen werden können, adressiert die systematische Literaturrecherche außerdem die Frage nach der Qualität verschiedener Imputationsverfahren. Daher werden Imputationsverfahren auf Grundlage der analysierten Vergleichsstudien dann auch auf qualitativer Ebene miteinander verglichen. Die zuvor herausgearbeiteten Vorgehensweisen können dabei als Hilfe dienen, um einen differenzierten Vergleich je Anwendungsfall anstellen zu können. Die Ergebnisse der systematischen Literaturrecherche werden im Anschluss an die Auswertung zum Abschluss des dritten Kapitels im Rahmen einer Diskussion zusammengefasst. Dabei werden sowohl die Vorgehensweisen zum Vergleich als auch die qualitative Bewertung der Imputationsverfahren noch einmal beleuchtet. Zusätzlich werden auch die betrachteten Veröffentlichungen an sich noch einmal kritisch beurteilt.

Im vierten Kapitel der Arbeit knüpft die exemplarische Anwendung verschiedener Imputationsverfahren im Kontext der Wissensentdeckung in Datenbanken an die systematische Literaturrecherche an. Dabei werden einige Erkenntnisse und möglicherweise identifizierte Forschungslücken der systematischen Literaturrecherche aufgegriffen. Dazu werden bestimmte Elemente der zuvor herausgearbeiteten Vorgehensweisen zum Vergleich von Imputationsverfahren in einen Wissensentdeckungsprozess in Datenbanken integriert. Folglich wird ein exemplarischer Vergleich von Imputationsverfahren als Vorbereitung für Data Mining anhand eines konkreten Fallbeispiels angestellt. Dabei werden anhand des Fallbeispiels die wesentlichen Schritte eines Wissensentdeckungsprozesses in Datenbanken beschrieben und durchgeführt, wobei die Imputationsverfahren auch dort den Kern der Ausführungen darstellen. Die Ergebnisse dieser exemplarischen Anwendung werden nach der Auswertung diskutiert und kritisch hinterfragt. Dabei werden die Ergebnisse auch in Bezug auf die Erkenntnisse der systematischen Literaturrecherche eingeordnet. Zu guter Letzt werden auf Grundlage der systematischen Literaturrecherche und der exemplarischen Anwendung Handlungsempfehlungen für die Auswahl geeigneter Imputationsverfahren im Rahmen von Wissensentdeckungsprozessen in Datenbanken abgeleitet.

Zum Abschluss dieser Arbeit werden diese Handlungsempfehlungen gemeinsam mit den wichtigsten Erkenntnissen und Schlussfolgerungen der gesamten Arbeit übersichtlich zusammengefasst. Letztlich schließt diese Arbeit mit einem kurzen Ausblick, in dem basierend auf den Ergebnissen dieser Arbeit Anstöße für zukünftige Forschungen im Bereich der Datenwissenschaft dargeboten werden.

2 Grundlagen der Datenanalyse

In diesem Kapitel werden alle notwendigen Grundlagen für die folgenden Ausführungen gelegt. Dazu werden zunächst die Zusammenhänge zwischen Daten, Information und Wissen erklärt, bevor speziell auf die Wissensentdeckung in Datenbanken eingegangen wird. In diesem Zuge wird auch der Begriff des Data Mining eingeführt, von der Wissensentdeckung abgegrenzt und in einen gemeinsamen Kontext gesetzt.

2.1 Zusammenhang zwischen Daten, Informationen und Wissen

Um ein Grundverständnis von Daten, Informationen und daraus hervorgehenden Wissensgewinnungsprozessen zu erlangen, ist zunächst eine Definition und Einordnung verschiedener Begriffe aus dem Bereich der Datenwissenschaft von Nöten. Dazu werden die Zusammenhänge zwischen Daten, Informationen und Wissen anhand der Wissenstreppe nach North erläutert, bevor allgemein in die Domäne der Datenwissenschaft eingeführt und Grundzüge der Datenorganisation erklärt werden.

North (2021) illustriert die Zusammenhänge zwischen Daten, Informationen und Wissen mithilfe der sogenannten Wissenstreppe. Die relevanten Grundelemente der Wissenstreppe nach North sind nachfolgend in Abbildung 2-1 dargestellt.

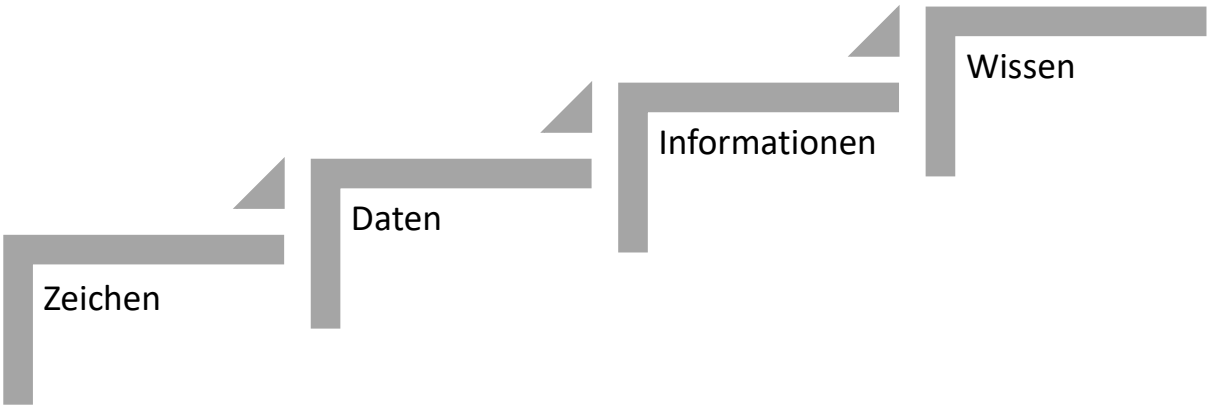

Abbildung 2-1: North'sche Wissenstreppe (eigene Darstellung in Anlehnung an North (2021, S. 37))

Die Wissenstreppe veranschaulicht, dass Daten lediglich mit Syntax versehene Zeichen sind. Demnach sind Datenelemente beliebige, nicht interpretierte Zeichen bzw. Zeichenfolgen. Wird diesen Daten nun eine Bedeutung zugeordnet, so entstehen Informationen. Werden diese Informationen weitergehend in einen Kontext gesetzt oder mit Erfahrungen kombiniert, so entsteht Wissen. North (2021, S. 37) beschreibt Wissen in diesem Zusammenhang als „Prozess der zweckdienlichen Vernetzung von Informationen durch das Bewusstsein". Demzufolge stellen Daten und daraus entstehende Informationen die Basis für Wissen dar. Probst, Raub &

Romhardt (2012) führen aus, dass Wissen alle zur Problemlösung eingesetzten Kenntnisse und Fähigkeiten umfassen. Sowohl diese Definition als auch die Wissenstreppe nach North implizieren, dass Wissen immer personengebunden ist. Die Nutzung von Wissen kann über weitere Stufen, die in der Abbildung nicht explizit abgebildet sind, schlussendlich zu Wettbewerbsfähigkeit bzw. zu Wettbewerbsvorteilen führen (North 2021).

Die Domäne der Datenwissenschaft beschäftigt sich mit ebendieser Ausschöpfung des Wissens und insbesondere mit der datengestützten Wissensgenerierung. Der englische Begriff Begriff *Data Science* wird seit den 1990er Jahren zunehmend genutzt und hat sich seitdem zu einer eigenständigen Wissenschaftsdisziplin entwickelt (Smith 2006). Laut Smith (2006, S. 163) umfasst die Datenwissenschaft demnach *„the study of the capture of data, their analysis, metadata, fast retrieval, archiving, exchange, mining to find unexpected knowledge and data relationships, visualization in two and three dimensions including movement, and management"*. Waller und Fawcett (2013, S. 78) beschreiben Datenwissenschaften als „Anwendung quantitativer und qualitativer Methoden, um relevante Probleme zu Lösen und Ergebnisse vorherzusagen" (eigene Übersetzung). Zusammenfassend ist die „Erfassung, Verarbeitung, Interpretation und Kommunikation von Daten mit dem Ziel der Gewinnung von Belastbarem und nutzenbringendem Wissen" (Plaue 2021, S. 1) die zentrale Aufgabe der Datenwissenschaft.

Grundlage für die Anwendung vieler Methoden aus dem Bereich der Datenwissenschaft ist ein Verständnis der Organisation von Daten. Dazu beschreiben Mertens et al. (2017) aufeinander aufbauende Begriffe der Datenorganisation folgendermaßen: Ein Datensatz fasst inhaltlich zusammenhängende Daten(-elemente) zusammen. Zusammengehörige Datensätze desselben Formats können weitergehend als Datei abgespeichert werden. Eine Datenbank wiederum bezeichnet die Sammlung zusammengehöriger Dateien auf Trägermedien. Datenbanken bilden zusammen mit Datenbankmanagementsystemen zur Verwaltung der Daten ein sogenanntes Datenbanksystem. Im Laufe der Zeit haben sich verschiedene Datenbankmodelle zur Speicherung von Datenstrukturen entwickelt. Während eine nähere Erklärung von Datentypen und -strukturen für das grundlegende Datenverständnis notwendig ist, ist eine umfangreiche Vorstellung von Datenbankmodellen für die folgenden Ausführungen nicht erforderlich.

Datentypen bezeichnen grundsätzlich die Zeichenart von Daten. Dabei lässt sich auf oberster Ebene nach numerischen (Ziffern), alphabetischen (Buchstaben) und alphanumerischen Daten (Ziffern, Buchstaben und Sonderzeichen) unterscheiden (Mertens et al. 2017). Darüber hinaus gibt es je nach Anwendung weitere Unterscheidungsmöglichkeiten für Datentypen.

Die Datenstrukturierung bzw. die Datenmodellierung ist essenziell zur Implementierung von Datenbanken (Mertens et al. 2017). Demnach ist die Aufgabe der Datenstrukturierung „die möglichst exakte Beschreibung des in der Datenbank abzubildenden Realitätsausschnittes"

(Mertens et al. 2017, S. 43). Das *Entity-Relationship-Modell* nach Chen (1976) hat sich als Standard zur Beschreibung von Datenstrukturen etabliert. Das Modell beschreibt Objekte und deren Beziehungen mithilfe von Attributen bzw. Merkmalen. Attribute können dabei sowohl die Objekt- als auch die Beziehungstypen näher beschreiben und werden dabei durch verschiedene Attributs- respektive Merkmalswerte definiert.

Objekte und deren Attribute werden in aller Regel bereits als Tabelle gespeichert oder lassen sich zumindest in tabellarische Form überführen. Dadurch werden Daten zu Anwendungszwecken zumeist in Form von Tabellen dargestellt. Dabei sind Tabellen bzw. Matrizen $(n \times m)$ Dimensionen, wobei n die Anzahl der Objekte und m die Anzahl der Attribute repräsentiert. Im Folgenden Verlauf dieser Arbeit werden diese Zusammenhängenden Datentabellen aus Objekten mit dazugehörigen Merkmalen als Datensätze bezeichnet.

2.2 Wissensentdeckung in Datenbanken und Data Mining

Nachdem der vorherige Abschnitt grundlegende Begriffe aus dem Feld der Datenwissenschaft erklärt und deren Zusammenhang darlegt, wird in den kommenden Ausführungen konkret auf den Prozess der Wissensentdeckung in Datenbanken eingegangen. Dazu wird zunächst der Begriff der Wissensentdeckung in Datenbanken definiert und vom Data-Mining-Begriff abgegrenzt. In diesem Rahmen werden auch erstmalig Data-Mining-Probleme bzw. -Aufgaben vorgestellt. Zudem wird der Begriff der Vorgehensmodelle eingeführt wird, ehe schließlich relevante Vorgehensmodelle präsentiert und miteinander verglichen werden.

Der Bereich der Datenanalyse lässt sich gemäß Knobloch und Weidner (2000) mithilfe zweier hauptsächlicher Analyseaufgaben unterteilen. Demnach stellt die datengetriebene Analyse mit dem Ziel der Mustererkennung einen Bottom-Up-Ansatz dar, wohingegen hypothesengetriebene Analyseaufgaben einen Top-Down-Ansatz darstellen. Während die Hypothesengetriebene Datenanalyse der Verifikation dient, verfolgt die Mustererkennung das Ziel, neues Wissen auf Basis von Daten zu generieren. Die Menge der Datenanalysemethoden zur Mustererkennung wird in der Literatur in der Regel als Data Mining bezeichnet (vgl. Knobloch und Weidner, 2000). Fayyad et al. (1996, S. 40) definieren Data Mining folgendermaßen:

> *„Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data."*

Diese Definition erklärt den Data-Mining-Begriff als die reine Anwendung hypothesenfreier Datenanalyseverfahren und führt gelichzeitig den Begriff des KDD-Prozesses ein. Data-Mining-Verfahren lassen sich dabei grundsätzlich nach Problem- bzw. Modellierungstyp differenzieren

(vgl. Jain und Srivastava 2013). Jain und Srivastava (2013) führen aus, dass deskriptive Modelle der Analyse hauptsächlicher Charakteristika der Daten selbst dienen, wohingegen vorhersagende Modelle gemäß dem Namen der Vorhersage bestimmter Variablen nützen. Diese beiden Modellierungstypen lassen sich wiederum durch untergeordnete Data-Mining-Aufgaben oder Analyseprobleme definieren. Eine mögliche Unterteilung unterteilt die deskriptiven Aufgaben in die Zusammenfassung, die Clusterbildung und die Assoziationsanalyse, wohingegen die Klassifizierung und die Regression den Vorhersageproblemen zugeordnet werden können (Gheware, Kejkar & Tondare 2014).

Da die beim Data Mining aufgedeckten Muster bzw. Modelle noch kein Wissen im Sinne der Wissenstreppe nach North darstellen, sind zusätzlich zum Data Mining vor- und nachgelagerte Schritte zur Gewinnung von Wissen notwendig. Der gesamte Prozess zur datengestützten Generierung von Wissen wird im Allgemeinen als *Knowledge Discovery in Databases (KDD)* bzw. als Wissensentdeckung in Datenbanken bezeichnet. Diese Bezeichnung wurde erstmals im Rahmen des *KDD Workshop 1989* verwendet und gewinnt seitdem im Umfeld der Datenwissenschaft zunehmend an Bedeutung. Fayyad et al. (1996) beschreiben KDD als „nichttrivialen Prozess der Identifizierung gültiger, neuer, potenziell nützlicher und schließlich verständlicher Muster in Daten“ (eigene Übersetzung). Basierend auf den Definitionen der Begriffe Data Mining und KDD nehmen Fayyad et al. (1996) folgende Einordnung vor:

> *“KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process.”*

Einige Veröffentlichungen setzen zwar Data Mining und KDD gleich (vgl. Hippner & Wilde 2001) oder fassen beide Begriffe als Knowledge Discovery and Data Mining (KDDM) zusammen (vgl. Asamoah & Sharda 2019), allerdings wird in dieser Arbeit eine ausdrückliche Trennung der Begriffe vorausgesetzt. Dabei wird Data Mining gemäß der zuvor aufgeführten Definition explizit als untergeordnete, wenn auch zentrale Phase eines KDD-Prozesses verstanden. Insgesamt besteht ein KDD-Prozess damit neben dem Data Mining aus zusätzlichen vor- und nachgelagerten Phasen (Knobloch und Weidner 2000).

2.2.1 Vorgehensmodelle zur Wissensentdeckung in Datenbanken

In den folgenden Abschnitten wird zunächst der Begriff des Vorgehensmodells eingeführt, bevor exemplarisch ausgewählte Vorgehensmodelle zur Wissensentdeckung in Datenbanken vorgestellt werden. Diese Vorstellung dient im Rahmen dieser Arbeit zum einen dem besseren Verständnis Wissensentdeckungsprozessen in Datenbanken. Zum anderen werden gleichzeitig die Datenvorverarbeitung und damit die Imputationsverfahren in den Gesamtkontext der Datenwissenschaft und der Wissensentdeckung in Datenbanken eingeordnet. Darüber hinaus

wird die exemplarische Anwendung im vierten Kapitel auf einem der vorgestellten Vorgehensmodelle aufbauen, weswegen besonders dieses Modell für diese Arbeit bedeutend ist.

Wie bereits im vorherigen Abschnitt ausgeführt, besteht der Wissensentdeckungsprozess in Datenbanken aus dem Data Mining sowie vor- und nachgelagerten Phasen. In der Vergangenheit haben sich darauf basierend sogenannte Vorgehensmodelle zur Wissensentdeckung in Datenbanken entwickelt. Ein Vorgehensmodell bezeichnet in diesem Kontext den Vorschlag eines Ablaufs gemäß festgelegten Schritten zum Zwecke der Wissensentdeckung in Datenbanken. Ein solches Modell besteht dabei in der Regel aus einer Reihe an Schritten bzw. Phasen, die nacheinander durchgeführt werden. Innerhalb dieser Phasen werden wiederum verschiedene Methoden bzw. Verfahren zur Zielerreichung angewendet.

Vorgehensmodell nach Fayyad

Fayyad et. al (1996) schlagen in den 1990er Jahren ein erstes Vorgehensmodell zur Wissensentdeckung in Datenbanken vor, das seitdem als Grundlage für Wissensentdeckungsprozesse und außerdem für weitere Vorgehensmodelle genutzt wird. Fayyad et al. (1996) greifen wesentliche Abläufe von Brachman und Anand (1996) auf und ergänzen diese zu einem Vorgehensmodell. Das eigentliche Vorgehensmodell besteht aus fünf datenbezogenen Ablaufphasen und wird von vier unterstützenden Schritten komplettiert, die auf der Interaktion des Anwenders basieren.

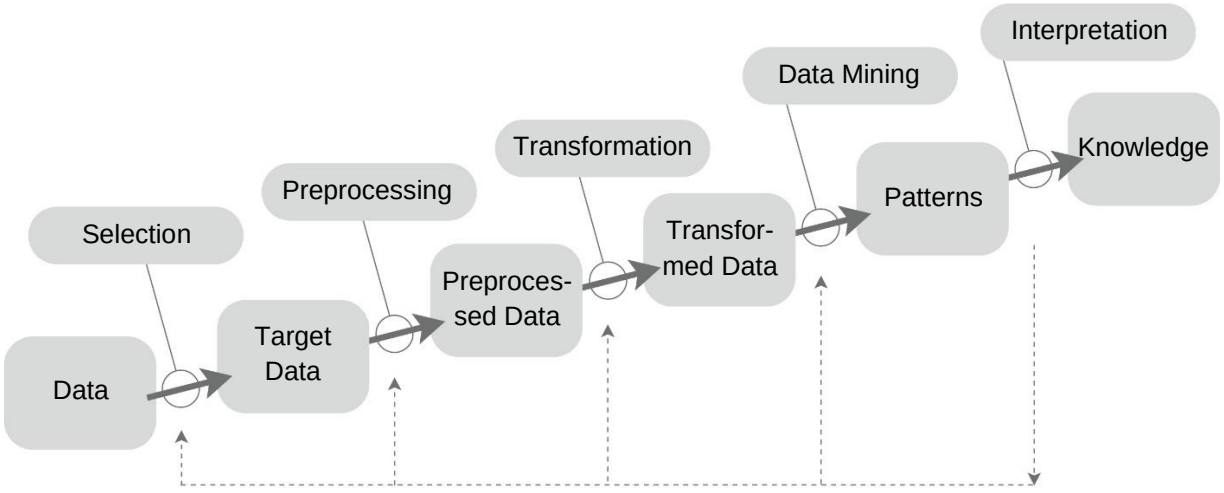

Abbildung 2-2: Vorgehensmodell nach Fayyad (eigene Darstellung in Anlehnung an Fayyad (1996, S. 3))

Im Folgenden werden anhand von Abbildung 2-2 die eigentlichen Phasen des Modells nach Fayyad (1996) erklärt, bevor anschließend noch einmal auf die unterstützenden aber nicht minder wichtigen Prozesse eingegangen wird:

(1) Datenauswahl: Bei der Datenauswahl wird ein geeigneter Datensatz ausgewählt, aus dem später neue Erkenntnisse gezogen werden sollen. Dieser Schritt kann außerdem die Beschränkung auf eine bestimmte Anzahl von Variablen bzw. die Auswahl von Stichproben enthalten. Die bei diesem Schritt ausgewählten Daten werden als Zieldaten bezeichnet.
(2) Datenvorverarbeitung: Dieser Schritt beschäftigt sich mit dem Umgang von Ausreißern und fehlenden Werten. Dazu wird ein Verständnis für die Datenunreinheiten aufgebaut, um die Daten gegebenenfalls davon zu bereinigen.
(3) Transformation: Hier werden die nun vorverarbeiteten Daten möglichst weit reduziert, ohne dabei Informationen zu verlieren oder andere Ergebnisse zu erhalten. Dazu werden beispielsweise nicht notwendige Attribute entfernt, um die Dimensionalität zu reduzieren.
(4) Data Mining: In diesem Schritt werden auf Grundlage der transformierten Daten schließlich Muster abgeleitet. Dazu werden je nach Data-Mining-Aufgabe unterschiedliche Methoden und Algorithmen angewendet
(5) Interpretation: In diesem abschließenden Schritt werden die Data-Mining-Ergebnisse mit dem Domänenwissen kombiniert, um schließlich neues Wissen zu generieren.
(Fayyad et al. 1996)

Das Vorgehensmodell ist dabei als iterativ anzusehen, wobei sowohl der gesamte Ablauf als auch einzelne Schritte wiederholt werden können. Weitergehend kann von jedem einzelnen Prozessschritt zu einem beliebigen vorangegangen Schritt (und nicht nur zum direkt vorgelagerten) gesprungen werden.

Die beschriebenen Abläufe und deren Ergebnisse sind maßgeblich von weiteren Begleitprozessen abhängig. Insbesondere für die Durchführung der ersten Phase, aber auch für die weiteren Phasen des Vorgehensmodells ist es von Bedeutung, dass zu Beginn ein umfangreiches Domänenwissen aufgebaut wird. Im Zuge dessen muss auch schon vor allen weiteren Phasen das Ziel des Data Mining definiert werden. Das hier definierte Ziel samt dem Verständnis der Domäne wird dann im Vorfeld des Data Mining wieder besonders wichtig. Hier wird anhand der Ziele zuerst die generelle Data-Mining-Methode ausgewählt, bevor in einem weiteren Schritt dann der zu verwendende Data-Mining-Algorithmus festgelegt wird. Im Anschluss an die Interpretation bzw. die Auswertung steht das Handeln auf Grundlage des neuen Wissens. Das Handeln kann dabei unterschiedlicher Natur sein und besteht zumindest aus der Dokumentation des neuen Wissens.

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

Der *Cross Industry Standard Process for Data Mining (CRISP-DM)* bietet einen standardisierten Prozess zur Umsetzung von Data-Mining- bzw. KDD-Prozesse (Chapman et al., 1999). CRISP-DM ist das Ergebnis eines Kooperationsprojektes verschiedener Unternehmen und wird an dieser Stelle vorgestellt, da es gleichzeitig ein frühes und noch immer häufig verwendetes Vorgehensmodell darstellt. Außerdem beinhaltet die ursprüngliche Veröffentlichung von Chapman et al. (1999) eine sehr detaillierte Anleitung auf mehreren Ebenen, die im Folgenden je nach Relevanz für diese Arbeit unterschiedlich ausführlich erläutert werden. Des Weiteren sind die Phasenbezeichnungen dieses Modells für spätere Ausführungen relevant, da diese in der später verwendeten Software *RapidMiner* zum Einsatz kommen.

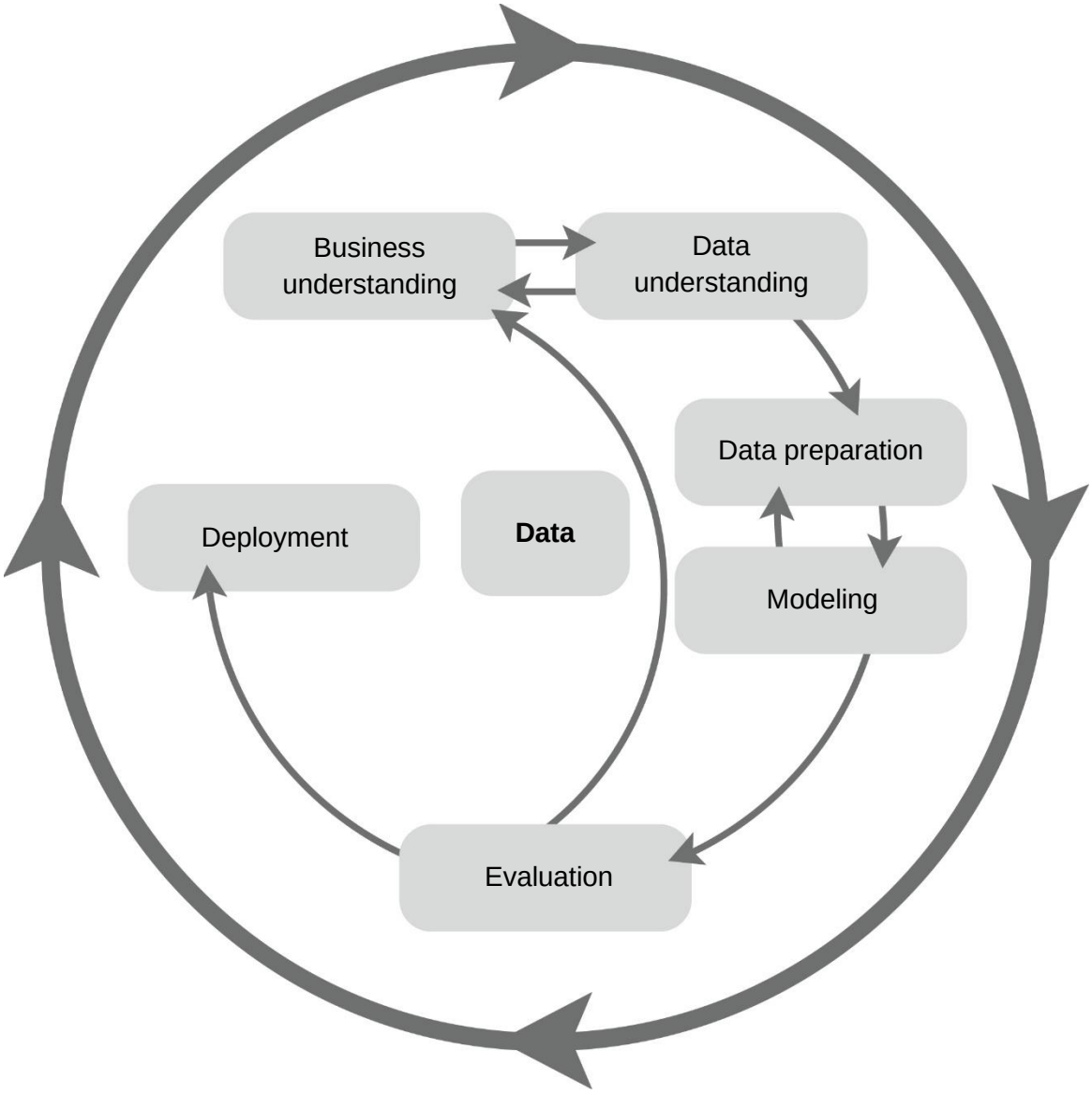

Abbildung 2-3: CRISP-DM-Vorgehensmodell (eigene Darstellung in Anlehnung an Chapman et al. (1999, S. 13))

CRISP-DM beschreibt auf vier Ebenen die Phasen, allgemeine Aufgaben, spezialisierte Aufgaben und Prozessbeispiele von KDD-Prozessen. Die erste Ebene enthält die allgemeinen Phasen, die mithilfe der weiteren Ebenen genauer spezifiziert werden. Da die Phasen die Grundzüge des Vorgehensmodells darstellen und am ehesten mit den Phasen der zuvor präsentierten Modellen vergleichbar sind, wird an dieser Stelle die oberste Ebene anhand von Abbildung 2-3 erklärt.

Die Abbildung illustriert den Lebenszyklus von Data-Mining- bzw. KDD-Prozessen. Dabei werden die wichtigsten Beziehungen und Wechselwirkungen zwischen den Phasen mithilfe von Pfeilen angedeutet, wobei nicht alle Beziehungen dargestellt werden können, da auch untergeordnete Aufgaben und Prozesse wiederum miteinander vernetzt sein können.

(1) *Business understanding* (Domänenverständnis) stellt in diesem Modell die erste Phase dar. Diese Phase beinhaltet das Aufbauen von Verständnis des Geschäftsfeldes und außerdem die Ableitung von Problemstellungen, Zielsetzungen und vorläufigen Umsetzungsplänen.
(2) *Data understanding* (Datenverständnis) beschreibt das Sammeln von Daten und darauf aufbauende Tätigkeiten zum Erlangen eines allgemeinen Datenverständnisses.
(3) *Data preparation* (Datenvorverarbeitung) fasst in diesem Modell alle Aktivitäten zur Auswahl und Vorbereitung des finalen Datensatzes zum Data Mining zusammen. Dieser Schritt umfasst damit nicht nur die Auswahl des Datensatzes und der relevanten Attribute, sondern auch die Transformation und Beseitigung von Datenunreinheiten.
(4) *Modeling* (Modellierung) befasst sich mit der eigentlichen Anwendung von Data Mining-Methoden. Da zumeist verschiedene Methoden mit jeweils verschiedenen Voraussetzungen zur Lösung von Data-Mining-Problemen existieren, kommen in dieser Phase besonders häufig Iterationen und Rückschritte in vorherige Phasen vor.
(5) *Evaluation* (Auswertung) bezeichnet sowohl die Bewertung der Modelle bzw. der Data-Mining-Ergebnisse als auch die der Durchführung der vorherigen Schritte selbst. Dazu wird beurteilt, ob die Ergebnisse die ursprünglich formulierten Ziele decken und ob die Ergebnisse im Folgenden in Geschäftsprozessen eingesetzt werden.
(6) *Deployment* (Umsetzung) ist die letzte Phase und damit der Abschluss dieses Modells. Hier werden die zuvor gewonnenen Ergebnisse bzw. das generierte Wissen in der Praxis eingesetzt. Die Spanne an Tätigkeiten reicht hier von der Dokumentation der Erkenntnisse bis hin zur Implementierung von veränderten oder neuen Geschäftsprozessen.

(Chapman et al. 1999)

Weitere relevante Vorgehensmodelle

Neben dem grundlegenden Vorgehensmodell nach Fayyad und dem deutlich detaillierteren CRISP-DM-Vorgehensmodell haben sich in der Literatur weitere, teilweise ähnliche Vorgehensmodelle herauskristallisiert. Zum anschließenden Vergleich und zur weiteren Einordnung der Vorgehensmodelle werden an dieser Stelle weitere relevante Vorgehensmodell verkürzt präsentiert, ohne dabei die genauen Phasen näher zu erklären.

So haben Hippner und Wilde (2001) ein domänenspezifisches Vorgehensmodell für die Anwendung im Bereich des Marketings entwickelt. Das Modell ist insgesamt in sieben Phasen unterteilt, die sowohl Aufgaben des Anwenders als auch rein datengestützte Methoden umfassen. Dabei sind mögliche Rückkopplungen, Wiederholungen und beliebiger Wechsel zwischen den Phasen ausdrücklich vorgesehen und Voraussetzung für zufriedenstellende Ergebnisse. Die Phasen der Aufgabendefinition (1), der Auswahl relevanter Datenbestände (2), der Datenaufbereitung (3), der Auswahl (4) und Anwendung (5) eines Data-Mining-Verfahrens sowie der Auswertung (6) und Anwendung (7) der Data-Mining-Ergebnisse orientieren sich dabei stark an dem Vorgehensmodell nach Fayyad. Das Vorgehensmodell nach Hippner und Wilde dient wiederum als Vorlage für das MESC-Modell, das den Ablauf auf die Domäne des Supply Chain Managements überträgt (Scheidler 2017). Das Modell verdeutlicht im Kontext dieser Arbeit einerseits, dass sich Abläufe zur Wissensentdeckung in Datenbanken je nach Domäne unterscheiden können. Andererseits zeigt die Tatsache, dass der Ablauf als Grundlage für ein Vorgehensmodell für eine andere Domäne dient, dass die Abläufe auch auf andere Domänen übertragen werden können.

Ein weiteres Vorgehensmodell namens Knowledge Discovery in Industrial Databases (KDID) der Autoren Lieber, Erohin und Deuse (2013) vereint gleich mehrere, vorherige Vorgehensmodelle und setzt diese in einen industriellen Kontext. Sie begründen, dass die gängigen Modelle meist für Branchen mit bereits verwendbaren Datenbeständen ausgelegt sind, was im industriellen Kontext allerdings nicht gegeben sei. Dadurch verdeutlicht das Vorgehensmodell für diese Arbeit abermals, dass eine Branchenspezifische Anpassung sowohl möglich als auch notwendig ist. KDID unterteilt den Gesamtablauf zur Wissensentdeckung im industriellen Kontext in insgesamt neun Phasen, darunter sechs KDD-typische und drei kontextbedingte Phasen. Darüber hinaus werden zwei Meilensteine definiert. Neben den KDD-typischen Schritten werden dabei die Aufnahme des Ist-Zustandes der IT-Struktur, die Datensammlung und -integration aus verschiedenen Quellen sowie die Entwicklung eines Werkzeugs zur Anwendung des generierten Wissens als kontextabhängige Phasen vorgeschlagen.

2.2.2 Vorgehensmodelle im Vergleich

In diesem Abschnitt werden die präsentierten Vorgehensmodelle abschließend miteinander verglichen, indem Unterschiede und besonders Gemeinsamkeiten der Modelle hervorzuheben. Zuletzt wird ein Vorgehensmodell ausgewählt, auf das sich die folgenden Ausführungen stützen werden. Dieser Abschnitt referenziert damit insbesondere die Einordnung der Datenvorverarbeitung und der Imputation fehlender Werte im Gesamtkontext der Wissensentdeckung in Datenbanken.

Wie bereits in der Überleitung zu diesem Kapitel dargelegt, besteht der Wissensentdeckungsprozess in Datenbanken auf der obersten Ebene aus dem Data Mining sowie vor- und nachgelagerten Ablaufphasen. Die hier vorgestellten Vorgehensmodelle zeigen, dass besonders die vor- und nachgelagerten Abläufe des Data Mining zwar in verschiedener Art und Weise, aber dennoch in ähnliche Phasen unterteilt werden. In den Grundzügen enthalten die Modelle dieselben Aufgaben, wobei diese teils unterschiedlich definiert oder verschiedenen Phasen zugeordnet werden. Zu Beginn des Prozesses enthalten alle Modelle den Verständnisaufbau für die jeweilige Domäne als explizite Phase (Hippner & Wilde, CRISP-DM, KDID) oder zumindest als unterstützenden Prozess (Fayyad). Ebenso stellt die Datenbasis in allen Modellen die Grundlage dar, wobei in einigen Modellen auch das Anlegen der Datenbasis als zusätzlicher Schritt definiert wird (KDID). Auf Grundlage der Datenbasis wird in allen Modellen im Vorfeld des Data Mining ein Datensatz selektiert, vorbereitet und transformiert. Dazu werden den Phasen jeweils ähnliche Aufgaben, wobei CRISP-DM all diese Schritte unter dem Vorbereitungsbegriff zusammenfasst. Während Data Mining in allen Modellen das zentrale Element des Prozesses darstellt, gibt es im Anschluss daran wieder kleine, definitorische Unterschiede in den Phasen. Im Grunde folgt in allen Modellen die Auswertung der Ergebnisse und zumindest in irgendeiner Form ein darauf basierendes Handeln.

Abseits der Definitionen der eigentlichen Ablaufschritte betonen die Autoren aller Modelle, dass einzelne Phasen zwar mithilfe von informationstechnischen Verfahren ablaufen, allerdings beruhen alle Modelle auf der Interaktion von Anwendern und sind zudem stark davon abhängig. Ebenso gehen alle Modelle in einem gewissen Maß auf die Wechselwirkungen zwischen den einzelnen Schritten ein und erlauben bzw. erfordern iteratives Vorgehen.

Bezüglich der Ablaufdefinition gibt es trotz der Vielzahl an Gemeinsamkeiten auch kleinere Unterschiede in der Beschreibung der Modelle. Das Vorgehensmodell nach Fayyad definiert Zwischenziele als Ergebnisse der Ablaufphasen. Im KDID-Prozess werden zwar ebenfalls Zwischenziele in Form von Meilensteinen definiert, allerdings werden diese nicht gesondert benannt, sondern lediglich durch die Durchführung der entsprechenden Schritte erreicht. CRISP-

DM definiert im Gegensatz dazu Zwischenziele als Ergebnisse der untergeordneten Aufgaben und generiert damit in jeder Phase diverse Ergebnisse.

Weitergehend lassen sich die Vorgehensmodelle nach dem Grad der Praxisorientierung unterscheiden. Während das Vorgehensmodell nach Fayyad einen sehr theoretischen Ansatz und damit eine Grundlage zur praktischen Anwendung liefert, stellt CRISP-DM durch die Beschreibung von Aufgaben und genaueren Abläufen ein praxisnäheres Modell dar. Sowohl das Modell nach Hippner und Wilde als auch KDID gehen noch einen Schritt weiter und beziehen sich auf jeweils eine spezielle Domäne. Das im Rahmen von Hippner und Wilde angesprochene MESC-Vorgehensmodell verdeutlicht die praktische Orientierung, indem es zeigt, dass sich die Grundzüge auch auf andere Geschäftsfelder übertragen lassen.

Die weiteren Ausführungen werden sich auf das CRISP-DM-Modell stützen, da es gleichzeitig auf den theoretischen Grundlagen nach Fayyad basiert und darüber hinaus eine detaillierte Anleitung zur praktischen Anwendung liefert. Diese Anleitung ist dabei durch die weitere Unterteilung der Phasen in Aufgaben, Aktivitäten und Ergebnisse gut strukturiert und eignet sich daher sowohl optimal zur Durchführung als auch zur Präsentation des Ablaufs. Außerdem verwendet die später eingesetzte Software teilweise die Bezeichnungen dieses Modells und ermöglicht damit eine direkte Umsetzung des Vorgehensmodells.

2.3 Datenvorverarbeitung

Um zu dem Umgang mit fehlenden Merkmalswerten und dem Füllen fehlender Daten überzuleiten, wird im Folgenden speziell die Phase der Datenvorverarbeitung thematisiert. Dazu werden die wesentlichen Aufgaben dieser Phase und deren Stellenwert noch einmal näher beleuchtet.

Die Datenvorverarbeitung ist gemäß dem vorherigen Vergleich eine essenzielle Phase aller Ablaufmodelle, der in der Regel auch dementsprechend viel Aufmerksamkeit zuteil wird. Da jeder KDD-Prozess individuell ist, ist nicht eindeutig zu bestimmen, wie groß der relative Zeitaufwand für die einzelnen Phasen ist. Allerdings begründen Kurgan und Musilek (2006) auf Grundlage mehrerer Studien, dass mindestens die Hälfte der Zeit eines jeden KDD-Prozesses für die Datenvorverarbeitung aufgewendet wird.

Da CRISP-DM im vorherigen Abschnitt als Verfahrensmodell für diese Arbeit festgelegt wurde und darüber hinaus die einzelnen Phasen mitsamt deren Aufgaben besonders detailliert beschreibt, wird die Datenvorverarbeitung nach CRISP-DM an dieser Stelle noch einmal näher beleuchtet. Dazu werden die definierten Aufgaben, beispielhafte Aktivitäten sowie deren Ergebnisse anhand von Abbildung 2-4 kurz präsentiert. Dabei steht die Datenbereinigung im Fokus, da diese den Umgang mit fehlenden Merkmalswerten umfasst.

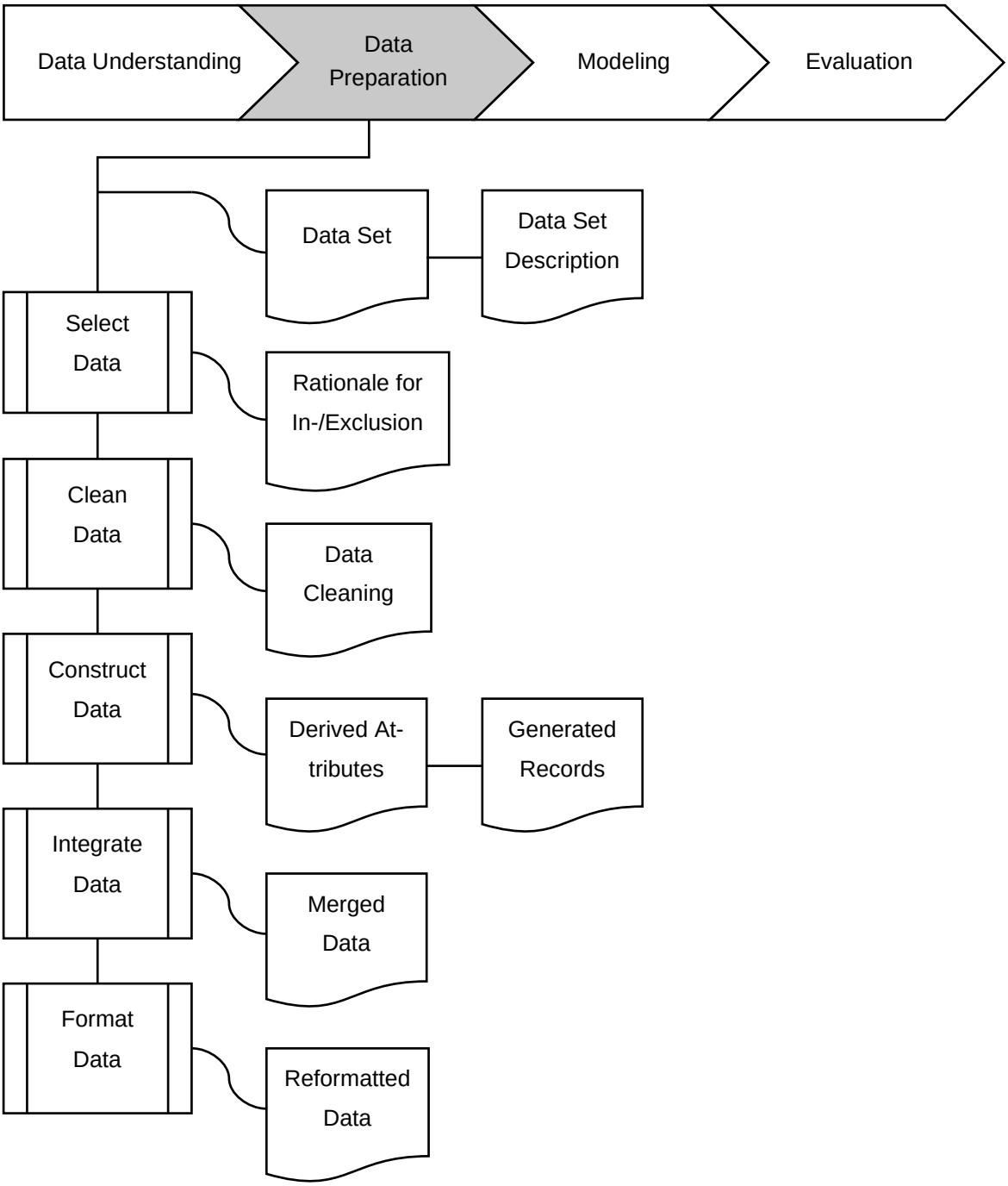

Abbildung 2-4: Aufgaben und Ergebnisse der Datenvorverarbeitung nach CRISP-DM (Eigene Abbildung in Anlehnung an Chapman et al. (1999, S. 23))

Die in der Abbildung aufgeführten Aufgaben, Aktivitäten und Ergebnisse werden in der Anleitung von Chapman et al. (1999) folgendermaßen definiert und beschrieben:

- Die Datenauswahl stellt zwangsläufig die erste Aufgabe der Datenvorverarbeitung dar, da alle weiteren Aufgaben und Aktivitäten auf Grundlage der ausgewählten Daten stattfinden. Innerhalb dieser Aufgabe werden sowohl eine möglichst repräsentative Daten-

menge als auch die zu betrachtenden Attribute ausgewählt. Das Ergebnis dieser Aufgabe wird durch eine Auflistung der verwendeten und nicht verwendeten Daten sowie einer Begründung für ebendiese Datenauswahl abgebildet. Die Besonderheit dieser Aufgabe ist, dass die hier getroffenen Entscheidungen im Anschluss an jede weitere Aufgabe der Datenvorverarbeitung evaluiert und daher gegebenenfalls wiederholt angepasst werden.

- Die Datenbereinigung stellt eine weitere und gleichzeitig essenzielle Aufgabe der Datenvorverarbeitung dar, da die Leistungsfähigkeit des Data Mining und damit auch die Ergebnisse des gesamten Vorgehens maßgeblich von der Qualität der bereinigten Daten abhängen. Daher wird in dieser Aufgabe die Qualität des ausgewählten Datensatzes angehoben und damit an die Ansprüche der ausgewählten Data-Mining-Techniken angepasst. Um diese Aufgabe zu erfüllen, werden Aktivitäten zum Umgang mit Rauschen festgelegt und angewendet. Rauschen ist nach CRISP-DM nicht genauer definiert, bezieht sich aber in aller Regel auf Ausreißer, fehlende Merkmalswerte oder besondere Werte. Hier gilt es, die Gründe für das Auftreten des Rauschens zu Verstehen und dementsprechend damit umzugehen. Dabei gilt es weitergehend festzustellen, welchen Einfluss das Rauschen auf das Data Mining hat und ob eine Bereinigung überhaupt sinnvoll und erforderlich ist. Mögliche Aktivitäten zum Bereinigen der Daten sind das Glätten des Wertebereichs oder das Füllen fehlender Merkmalswerte. Das Ein- oder Ersetzen von Werten kann dabei verschieden aufwendig realisiert werden, indem Standartwerte eingesetzt oder plausible Werte durch Modellierungsverfahren abgeschätzt werden. Das Ergebnis dieser Aufgabe ist ein Bericht über die Durchführung aller Aktivitäten mitsamt der jeweiligen Begründung.
- Je nach Anwendungsfall kann außerdem die Datenkonstruktion notwendig oder hilfreich sein. Dabei ist es das Ziel, neue und für das weitere Vorgehen hilfreiche Attribute zu erzeugen oder die Werte existierender Attribute zu transformieren. Das Ergebnis dieser Aufgabe sind dementsprechend abgeleitete Attribute und generierte Daten.
- Ein weiterer Bestandteil der Datenvorverarbeitung ist die Datenintegration. Hier werden Daten verschiedener Datensätze miteinander kombiniert oder externe Informationen zugeführt, um neue Daten oder Werte zu erzeugen. Dazu können Werte gänzlich neu hinzugefügt oder aus mehreren Werten zusammengesetzt werden. Das Ergebnis dieser Aufgabe ist ein zusammengeführter Datensatz.
- Die abschließende Aufgabe der Datenvorverarbeitung besteht in der Datenformatierung. Diese Aufgabe beschreibt die Transformation der Daten auf syntaktischer Ebene,

denn je nach Data-Mining-Technik und -Werkzeug existieren verschiedene Anforderungen an Datentypen oder Attributsreihenfolge. Das Ergebnis dieser Aufgabe wird dementsprechend durch einen formatierten Datensatz dargestellt.

- Das Gesamtergebnis der Datenvorverarbeitung ist folglich der finale Datensatz, der im weiteren Prozess zur Modellierung bzw. zum Data Mining eingesetzt wird. Dieser Datensatz wird dabei durch eine Beschreibung ergänzt, die dem näheren Verständnis der Daten dient.
  (Chapman et al. 1999)

2.4 Umgang mit fehlenden Merkmalswerten

Nachdem zu Beginn des zweiten Kapitels dieser Arbeit in die Thematik der Wissensentdeckung in Datenbanken und später in den Bereich der Datenvorverarbeitung eingeführt wurde, beschäftigen sich die folgenden Abschnitte speziell mit dem Umgang von fehlenden Merkmalswerten. Gemäß dem in dieser Arbeit angewendeten CRISP-DM-Vorgehensmodell stellt der Umgang mit fehlenden Merkmalswerten eine Aktivität der Datenbereinigung dar und ist damit Teil der Datenvorverarbeitung. Dazu werden wichtige Begrifflichkeiten zu fehlenden Merkmalswerte kurz erläutert, bevor zu den Verfahren zur Behandlung fehlender Werte übergeleitet wird.

2.4.1 Fehlende Merkmalswerte

In der Realität sind fehlende Daten keine Ausnahme, sondern die Regel. Je nach Ursprungsdomäne der Daten gibt es kaum bis keine vollständig gefüllten Datensätze. Das Auftreten fehlender Werte innerhalb eines Datensatzes kann dabei anhand verschiedener Merkmale, namentlich dem Ausfallmuster, der Ausfallrate und dem Ausfallmechanismus, beschrieben werden (vgl. Rockel 2017; Lang & Little 2018). Diese Merkmale fehlender Daten sind im Kontext der Datenbereinigung relevant, da die Verfahrensauswahl zur Behandlung der fehlenden Werte sowie die Ergebnisse mitunter davon abhängen können.

Ausfallmuster beschreiben in diesem Zusammenhang, ob Werte eines oder mehrerer Merkmale fehlen. Fehlen nur Werte eines Merkmals, so spricht man von einem univariaten Ausfallmuster, während das Fehlen von Werten mehrerer Merkmale als multivariates Ausfallmuster bezeichnet wird (van Buuren 2012).

Die Ausfallrate beschreibt gemäß dem Namen den Anteil der fehlenden Werte innerhalb eines Datensatzes. Die Ausfallrate schließt zwar per se keine Verfahren zur Behandlung fehlender Werte aus, allerdings können die Ergebnisse der Verfahren je nach Ausfallrate unterscheiden.

Der Ausfallmechanismus unterscheidet nach der Ursache für das Auftreten fehlender Werte und hat den wohl größten Einfluss auf die Verfahrensauswahl.

Bei den Ausfallmechanismen kann auf oberster Ebene zwischen dem systematischen bzw. nicht zufälligen und dem unsystematischen bzw. zufälligen Auftreten fehlender Daten unterschieden werden (Bankhofer 1995). Eine detailliertere und wohl am weitesten verbreitete Unterteilung unterscheidet folgende Ausfallmechanismen: *missing completely at random (MCAR), missing at random (MAR)* und *missing not at random (MNAR)* (Little & Rubin 2020). MCAR beschreibt gemäß der Bezeichnung das komplett zufällige Auftreten fehlender Daten. Fehlende Merkmalswerte können bei diesem Ausfallmechanismus nicht durch Relationen zu anderen Merkmalen oder deren Ausprägungen erklärt werden. MAR bezeichnet dementsprechend den Fall, dass sich fehlende Werte mithilfe der Beziehung zu anderen Merkmalswerten bzw. deren Ausprägungen erklären lassen. MNAR beschreibt weitergehend das objektbedingte Fehlen von Werten. MNAR tritt damit in Situationen auf, in denen zwar ein Grund für das Fehlen von Werten vorliegt, dieser jedoch nicht durch den Datensatz abgebildet wird.

2.4.2 Verfahren zum Umgang mit fehlenden Merkmalswerten

Nachdem die vorherigen Abschnitte das Auftreten fehlender Daten und die deren Behandlung thematisieren, werden an dieser Stelle die Möglichkeiten zum Umgang mit den fehlenden Daten erörtert. Wie im Vorfeld dieses Kapitels angemerkt, ist es für KDD-Prozesse erforderlich, den Umgang mit fehlenden Merkmalswerten zu thematisieren, da eine Fortsetzung weiterer KDD-Phasen, insbesondere dem Data Mining, mit unvollständigen Daten nicht unmittelbar möglich ist.

Bankhofer (1995) unterteilt die Verfahren zum Umgang mit fehlenden Daten auf oberster Unterscheidungsebene in vier Gruppen, die in Abbildung 2-5 dargestellt sind. Dabei lässt ich jede einzelne Gruppe weitergehend unterteilen und besteht letztendlich aus einer Sammlung verschiedener Techniken zum Umgang mit fehlenden Merkmalswerten. Bankhofer erklärt eine Vielzahl dieser Techniken in seinem Grundlagenwerk vergleichsweise detailliert, ohne dabei jedoch einzelne Softwareimplementationen oder spezielle Algorithmen zu beschreiben, weswegen sich die folgenden Ausführungen immer wieder auf Bankhofers Ausführungen beziehen.

Die Anpassung oder Auswahl der anschließenden Data-Mining-Verfahren, sodass diese auch mit fehlenden Daten durchgeführt werden können, ist gesondert aufgeführt, da der Datensatz hierbei unberührt bleibt. Bankhofer (1995) bezeichnet diese Strategie auch als multivariate Datenanalyse. Die verbleibenden drei Gruppen werden im Gegensatz dazu durch Verfahren gebildet, bei denen der zugrundeliegende Datensatz verändert respektive ergänzt wird. An

dieser Stelle werden die grundlegenden Merkmale dieser Gruppen zur Abgrenzung voneinander kurz präsentiert, bevor in den folgenden Kapiteln dann ausführlich auf Imputationsverfahren eingegangen wird.

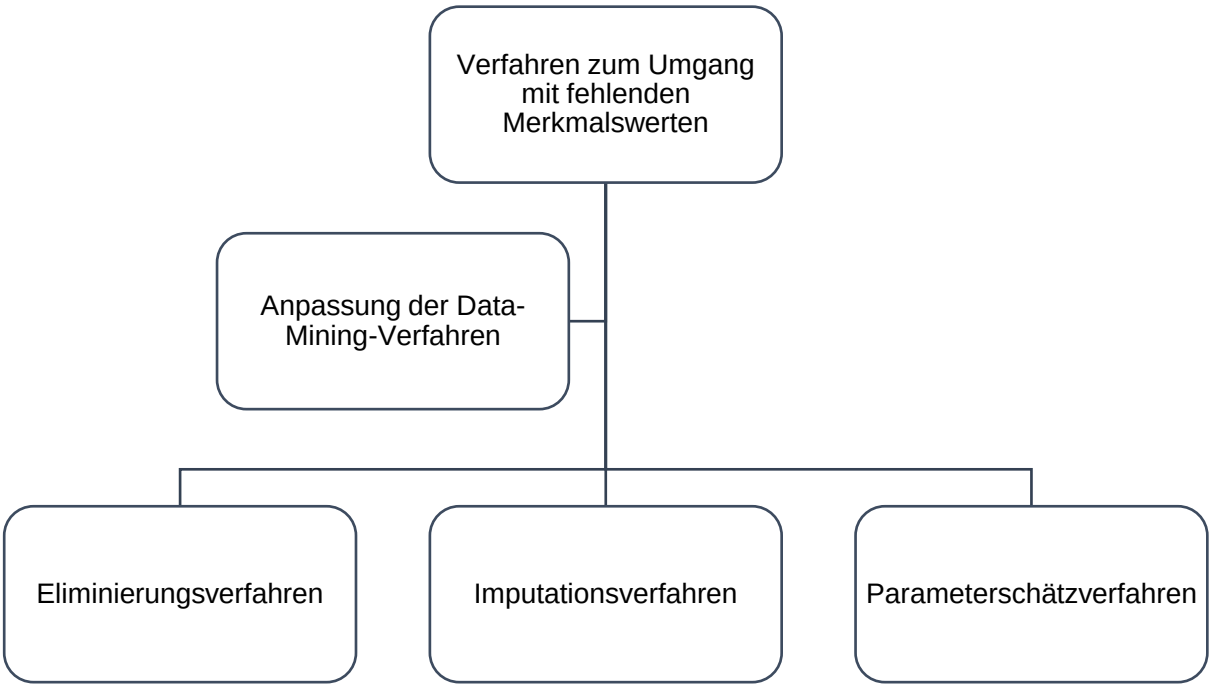

Abbildung 2-5: Einteilung von Verfahren zum Umgang mit fehlenden Werten (eigene Darstellung in Anlehnung an Bankhofer (1995, S. 89))

Eliminierungsverfahren stellen eine triviale Möglichkeit dar, Datensätze von fehlenden Werten zu bereinigen. Prinzipiell werden bei Eliminierungsverfahren entweder Objekte oder Merkmale mit fehlenden Werten aus der Datenmatrix entfernt. Dadurch lassen sich diese Verfahren in Objekt- und Merkmalseliminierung unterscheiden. Weitergehend existieren zwei verschiedene Ansätze zur weiteren Datenanalyse, die auf Eliminierungsverfahren beruhen. Einerseits besteht die Möglichkeit zur Analyse aller vollständigen Objekte bzw. Merkmale, indem alle Zeilen bzw. Spalten mit fehlenden Merkmalswerten gänzlich entfernt werden. Diese Strategie bezeichnet Bankhofer als *complete-case analysis*. Andererseits besteht die Möglichkeit zur Auswertung der verfügbaren Objekte, die analog als *available-case analysis* bezeichnet wird. Für die Objekteliminierung bedeutet das, dass univariate Statistiken für diejenigen Objekte erhoben werden, bei denen das betrachtete Merkmal tatsächlich ausgeprägt ist. Dadurch variiert die Stichprobengröße je nach betrachtetem Merkmal. Dieses Prinzip lässt sich analog auch auf die Merkmalseliminierung übertragen.

Imputationsverfahren stellen die wohl wichtigste Gruppe zum Umgang mit fehlenden Merkmalswerten dar, da in nahezu jedem Wissensentdeckungsprozess mit unvollständigen Daten

eines dieser Verfahren angewendet wird. Imputationsverfahren werden im Deutschen auch als Ersetzungs-, Ergänzungs- oder Vervollständigungsverfahren bezeichnet (Bankhofer 1995), wohingegen im Englischen zumeist der analoge Begriff *imputation methods* verwendet wird. Grundsätzlich werden die fehlenden Werte mithilfe ausgewählter Techniken auf Grundlage der vorhandenen Daten geschätzt, wodurch folglich eine vollständige Datenmatrix entsteht.

Bezüglich der Parameterschätzverfahren bezieht sich Bankhofer (1995) auf eine Definition, die darunter alle Methoden zum Schätzen fehlender Werte zusammenfasst. Da hierbei jedoch keine Abgrenzung zu Imputationsverfahren möglich ist, ordnet Bankhofer in seinen Ausführungen den Parameterschätzverfahren nur diejenigen Verfahren zu, bei denen die Schätzungen nicht auf den vollständig vorliegenden Objekten beruhen. Für diese Arbeit sind Parameterschätzverfahren vorerst nicht weiter von Bedeutung, sondern verdeutlichen lediglich, dass neben der Imputation auf Grundlage der vorhandenen Daten noch weitere Schätzverfahren existieren.

2.5 Imputationsverfahren

Nachdem die unmittelbar vorausgehenden Ausführungen in den Umgang mit fehlenden Merkmalswerten eingeführt haben, widmen sich die folgenden Ausführungen nun konkret den Imputationsverfahren. Da im weiteren Verlauf dieser Arbeit sowohl ein Vergleich auf Grundlage der bisherigen Literatur als auch anhand eines exemplarischen Anwendungsbeispiels vorgenommen wird, werden in diesem Kapitel die dazu notwendigen Grundlagen gelegt. Dazu wird zunächst an den Abschnitt zu fehlenden Merkmalswerten angeknüpft, indem Imputationsverfahren noch einmal näher definiert und eingeordnet werden. Dazu wird auf die allgemeinen Vor- und Nachteile, die Ziele und die Unterscheidungsmöglichkeiten von Imputationsverfahren eingegangen. Anschließend werden die gängigsten Verfahren bzw. Verfahrensgruppen in den Abschnitten 2.5.1 und 2.5.2 kurz vorgestellt.

Wie in Kapitel 2.4.2 dargelegt, umfassen Imputationsverfahren nach Bankhofer (1995) alle Schätzverfahren zum Füllen fehlender Werte auf Grundlage der vorhandenen Werte eines Datensatzes. In dieser Arbeit werden alle Verfahren , mithilfe derer fehlende Daten ausgefüllt werden, als Imputationsverfahren verstanden.

Allgemeine Vor- und Nachteile von Imputationsverfahren

Schafer und Graham (2002) führen gleich mehrere allgemeine Vorteile von Imputationsverfahren auf. Demnach besteht ein großer Vorteil von Imputationsverfahren im Allgemeinen darin, dass keine Objekte entfernt werden und damit alle beobachteten Daten für anschließende

Analysen verwendet werden. Darüber hinaus vereinfachen Imputationsverfahren die weitere Verwendung zu Analyse- bzw. Data-Mining-Zwecken, da Analysewerkzeuge in der Regel für vollständige Datensätze ausgelegt sind. Außerdem erklären die Autoren, dass bei entsprechender Datengrundlage sehr präzise bzw. plausible und damit gut verwendbare Werte imputiert werden können, die gegebenenfalls verborgene Informationen enthalten.

Auf der anderen Seite führen Schafer und Graham (2002) aus, dass die Imputation und deren Anwendung auch einige Gefahren birgt. Imputationsverfahren bedürfen demzufolge einiger Vorbereitung und können, insbesondere bei multivariaten Daten, schwierig anzuwenden sein. Zudem können insbesondere einfache Verfahren das statistische Verhalten und die Beziehungen zwischen einzelnen Daten verzerren. Little und Rubin (2020) konkretisieren, dass wichtige statistische Kennzahlen der Daten, wie z. B. der Mittelwert, die Standardabweichung und die Varianz, durch Imputationsverfahren verändert werden können. Derartige Verzerrungen treten insbesondere bei systematischem Fehlen von Werten (MNAR) auf und können damit auch verzerrte oder gar falsche Analyseergebnisse zur Folge haben (Huismann 2000).

Ziele von Imputationsverfahren

Um den gerade dargelegten Gefahren vorzubeugen und die Vorteile bestmöglich auszunutzen, formuliert Sande (1982) folgende Ziele bzw. Ansprüche, die an Imputationsverfahren gestellt werden. Demnach sollen Imputationsverfahren (1) plausible Werte imputieren, (2) Verzerrungen reduzieren und die Beziehungen innerhalb der Daten bestmöglich erhalten, (3) auf möglichst alle Ausfallmuster anwendbar sein, (4) im Voraus eingerichtet werden können und (5) in Bezug auf Verzerrungen und Vorhersagegenauigkeit evaluierbar sein.

Unterscheidung von Imputationsverfahren

In der bisherigen Literatur zum Umgang mit fehlenden Merkmalswerten haben sich verschiedene Möglichkeiten zur Einteilung von Imputationsverfahren herausgebildet. Einige gängige Einteilungsmöglichkeiten werden im folgenden Abschnitt vorgestellt, um die einzelnen Methoden anschließend strukturiert vorstellen zu können.

Nordholt (1998) unterscheidet auf oberster Ebene in deduktive, deterministische und stochastische Methoden. Deduktive Methoden umfassen dabei diejenigen Techniken, bei denen die fehlenden Merkmalsausprägungen direkt anhand der beobachteten Ausprägungen desselben Objektes bestimmt werden können. Die Gruppen der stochastischen und deterministischen Methoden werden hingegen durch diejenigen Methoden gebildet, die fehlende Merkmalswerte auf Grundlage der beobachteten Werte abschätzen. Nach der zuvor verwendeten Definition sind damit insbesondere die stochastischen und deterministischen Methoden der Betrach-

tungsgegenstand dieser Arbeit. Deterministische Verfahren lassen sich nach dieser Unterteilung dadurch charakterisieren, dass die bei Wiederholung dieselben Imputationswerte liefern, wohingegen stochastische Verfahren irgendeine Art von Zufallsprozessen verwenden (Nordholt 1998; Huismann 2000).

Das bereits zuvor herangezogene Grundlagenwerk von Bankhofer (1995) unterscheidet Imputationsverfahren hauptsächlich in drei Gruppen, namentlich in einfache Imputationstechniken, Imputation innerhalb von Klassen und multivariate Imputationstechniken. Dabei werden Parameterschätzverfahren gesondert und damit nicht als Teil der Imputationsverfahren aufgeführt.

In neueren Veröffentlichungen werden Imputationsverfahren zumeist in zwei übergeordnete Gruppen eingeteilt, bevor die einzelnen Methodengruppen und Techniken selbst näher charakterisiert werden. Baraldi und Enders (2010) benennen diese Gruppen als s*ingle imputation methods* und *modern missing data techniques*. Demnach umfassen single imputation methods die traditionellen Verfahren zum Ausfüllen der fehlenden Werte durch eine Vorhersage auf Grundlage der beobachteten bzw. vorhandenen Daten. Modern missing data techniques werden nach dieser Definition durch die ‘state of the art‘ missing data techniques repräsentiert. Die Modernen Imputationsmethoden sind den traditionellen Methoden nach dieser Definition aufgrund der Tatsache überlegen, dass die sowohl bei MCAR- als auch bei MAR-Daten unverzerrte Schätzungen versprechen. Eine ähnliche Definition unterscheidet in *Single-Value Imputation* und *Model-Based Imputation* (MIT Critical Data 2016). Single-Value-Imputation-Methoden beschreiben nach dieser Definition das einmalige Ausfüllen durch einen vorhergesagten Wert. Modellbasierte Methoden beschreiben hingegen all diejenigen Verfahren, bei denen die fehlenden Werte auf Grundlage eines Vorhersagemodells geschätzt werden, das anhand der vorhandenen Werte entwickelt wird. Diese Bezeichnung bzw. Definition konkretisiert zwar insbesondere die zweite Gruppe, allerdings wird eine genaue Einteilung einzelner Verfahren dadurch erschwert. Moderne Verfahren, die nicht auf Vorhersagemodellen basieren, ließen sich nach dieser Einteilung in keine der beiden Gruppen einordnen. Außerdem widerspricht diese Einteilung der Unterscheidung nach Little und Rubin (2020), die ausführen, dass sich auch Single-Imputation-Methoden als explizite und implizite Modellierungsmethoden beschreiben lassen.

Eine weitere in der Literatur eher selten zu findende aber dennoch für diese Arbeit wichtige Einteilung unterscheidet nach Imputationsverfahren für longitudinale und nicht-longitudinale Daten (vgl. Kaya & Turkoglu 2021). Zwar lassen viele der Methoden auf longitudinalen und nicht-longitudinalen Datensätze anwenden, allerdings verschwimmen insbesondere bei der Imputation für longitudinale Daten die Grenzen zwischen Imputation und Data Mining. Werden

bei longitudinalen Datensätzen nicht nur bestimmte fehlende Merkmalsausprägungen, sondern auch die Zielvariable der Datenanalyse mit Regressionstechniken imputiert, so dient die Imputation gleichzeitig der Datenanalyse und findet damit nicht mehr im Kontext der Datenvorverarbeitung statt. Darüber hinaus existieren spezielle Methoden, die die zeitliche Abhängigkeit longitudinaler Daten berücksichtigen und damit nur bei longitudinalen Datensätzen anwendbar sind.

Um die relevanten Imputationsmethoden strukturiert vorzustellen, wird in dieser Arbeit gemäß der Definition von Baraldi und Enders (2010) eine Unterscheidung in traditionelle und moderne Verfahren vorgenommen. Dabei werden die Imputationsmethoden bzw. übergeordnete Methodengruppen allgemein vorgestellt, ohne dabei näher auf die konkreten statistischen und mathematischen Methoden einzugehen. Der Vollständigkeit wegen muss auch hier angemerkt werden, dass eine eindeutige, objektive Zuordnung nicht immer möglich ist und dass einzelne Techniken gegebenenfalls nicht erfasst werden.

2.5.1 Traditionelle Imputationsmethoden

Wie im vorherigen Abschnitt dargelegt, werden traditionelle Imputationsmethoden in dieser Arbeit als diejenigen Methoden verstanden, die fehlende Werte ohne großen Aufwand durch einmaliges Schätzen auf Grundlage der vorhandenen Werte ausfüllen. In diesem Abschnitt werden in der bisherigen Literatur beschriebene Methoden benannt, wobei die gängigsten dieser Methoden zum Verständnis der folgenden Kapitel noch einmal näher charakterisiert.

Mittelwert-, Median- und Modalwertsimputation

Die Ausprägungen eines Merkmals können nach Bankhofer (1995) durch sogenannte Lageparameter beschrieben werden, die als mögliche Imputationswerte für fehlende Werte verwendet werden können. Nach Bankhofer eignen sich das arithmetische Mittel, der Median und der Modalwert als verwendbare Lageparameter. Zur Imputation wird demzufolge der entsprechende Lageparameter anhand aller vorhandenen Merkmalswerte bestimmt und für alle fehlenden Werte eingesetzt. Dabei ist anzumerken, dass Median und Modalwert vergleichsweise selten zur Imputation genutzt werden, wohingegen die Imputation des arithmetischen Mittels die älteste und eine häufig verwendete Methode darstellt. Die Mittelwertsimputation ist in der Literatur zumeist unter dem englischen Begriff *Mean Imputation* (vgl. Little & Rubin 2020; Baraldi & Enders 2010) oder *Mean Substitution* (vgl. Schafer & Graham 2002) zu finden. Little und Rubin (2020) unterscheiden die Mittelwertsimputation dabei zusätzlich in die „klassische“, bedingungslose Mittelwertsimputation und die fortgeschrittene, bedingte Mittelwertsimputation. Dabei bezieht sich die bedingungslose Mittelwertsimputation auf das zuvor beschriebene

Einsetzen des Mittelwerts aller vorhandenen Ausprägungen. Die bedingte Mittelwertsimputation kann hingegen durch verschiedene Anpassungen realisiert werden. Little und Rubin (2020) führen beispielsweise das Bilden von Klassen auf Grundlage der beobachteten Merkmale und die anschließende Mittelwertsimputation innerhalb der jeweiligen Klasse auf. Diese Technik lässt sich analog dazu auch auf Imputation mithilfe anderer Lageparameter übertragen, was Bankhofer (1995) als Imputation des Klassenlageparameters bezeichnet.

Zufallszahlimputation

Die Zufallszahlimputation ist eine triviale und in der Praxis selten verwendete Methode zur Vervollständigung von Datenmatrizen. Gemäß dem Namen werden bei Techniken der Zufallszahlimputation zufällige Zahlen eines bestimmten Wertebereichs für anstelle der fehlenden Werte ergänzt (Bankhofer 1995). Dabei unterscheidet Bankhofer (1995) grundlegend in zwei verschiedene Techniken, um den Wertebereich für den zu ergänzenden Wert zu bestimmen: Einerseits kann der Wertebereich komplett willkürlich unter Beachtung des zu ergänzenden Datentyps gewählt werden, wodurch vollständig zufällige Werte generiert und ergänzt werden. Andererseits kann der Wertebereich für die jeweils zu ergänzende Merkmalausprägung durch die Menge der beobachteten Ausprägungen gebildet werden.

Cold-Deck und Hot-Deck Methoden

Bei den sogenannten Deck-Methoden werden die fehlenden Ausprägungen unvollständiger Objekte grundsätzlich durch den Wert eines Spenderobjektes ergänzt (vgl. Nordholt 1998). Dabei lässt sich weitergehend zwischen den sogenannten Cold- und Hot-Deck Methoden unterscheiden, wobei auch diese sich wiederum in untergeordnete Techniken gliedern lassen.

Für die Imputation mittels der Cold-Deck Methoden werden die Spenderobjekte und Werte aus einem weiteren, externen Datensatz bezogen. Um geeignete Spenderobjekte zu identifizieren, werden im Vorfeld der Imputation bestimmte Kovariaten bzw. Entsprechungsschlüssel (Nordholt 1998) oder Imputationsklassen (Bankhofer 1995) festgelegt, die bei Spender- und Empfängerobjekt übereinstimmen oder sich zumindest ähneln müssen. Bankhofer (1995) führt zu Cold-Deck Methoden jedoch aus, dass diese in der Praxis nur in seltenen Fällen eingesetzt werden können, da zusätzlich zum eigentlichen Datensatz ein weiterer Datensatz mit denselben Merkmalen von Nöten ist.

Für die Imputation mittels der Hot-Deck Methode werden Ausprägungen bzw. Spenderobjekte aus dem Datensatz selbst verwendet. Nordholt (1998) unterscheidet dabei in zwei Techniken: Wird das vorangegangene Objekt eines unvollständigen Objektes als Spenderobjekt verwendet, so wird von der sequenziellen Hot-Deck Methode gesprochen. Wird ein zufälliges Objekt

des Datensatzes als Spenderobjekt ausgewählt, wird das als zufällige Hot-Deck Methode bezeichnet. Die zufällige Hot-Deck Methode ist damit eine Form der zuvor beschriebenen Zufallszahlimputation. Bankhofer (1995) führt zusätzlich wieder die Imputation einer Ausprägung eines Objektes derselben Imputationsklasse auf, um möglichst plausible Werte zu ergänzen.

Regressionsimputation

Die Regressionsimputation stellte ursprünglich neben der Mittelwertsimputation die am weitesten verbreitete Gruppe von Imputationsmethoden dar (Bankhofer 1995), wodurch es sich um eine traditionelle und grundlegende Form der Imputation handelt. Weitergehend handelt es sich bei der Regressionsimputation um eine erste Form der modellbasierten Imputationsmethoden, bei der fehlende Werte unter Verwendung von Regressionsmodellen ergänzt werden (MIT Critical Data 2016). Dazu wird prinzipiell ein Regressionsmodell anhand einer oder mehrerer ausgewählter Variablen aufgestellt und auf den Datensatz angewendet, sodass die Werte einer Zielvariable ergänzt werden können (Zhang 2016). Grundsätzlich lassen sich unter dem Begriff der Regressionsimputation demnach alle Imputationsverfahren zusammenfassen, die Werte auf Grundlage einer Regressionsanalyse ergänzen. Weitergehend ist anzumerken, dass in der Literatur eine Vielzahl von Ausprägungsformen der Regressionsimputation, wie zum Beispiel die stochastische Regression, die lineare Regression oder die logistische Regression, zu finden sind, an dieser Stelle aber nicht gesondert ausgeführt werden.

Zwar bezeichnen Lang und Little (2018) die einfache Regressionsimputation als altmodische Imputationsmethode, allerdings werden Regressionsmodelle noch immer für moderne Verfahren als Grundlage verwendet, wodurch der Regression im Bereich der Imputation eine noch immer große Bedeutung zukommt. Während Bankhofer im Jahr 1995 noch ausführte, dass zur Imputation fast ausschließlich lineare Regressionsmodelle zur Anwendung kommen, hat sich in den letzten Jahren ein Trend hin zu nichtlinearen, adaptiven Regressionsmodellen entwickelt (vgl. Sanchez Lasheras et al. 2020; Jove et al. 2018; Crespo Turrado et al. 2015). Außerdem lassen sich Regressionsmodelle als Grundlage für multiple Imputationsverfahren, die im Nachgang noch näher thematisiert werden, verwenden (vgl. van Buuren & Groothuis-Oudshoorn 2011). Die Verwendung nichtlinearer Regressionsmodelle oder die Verwendung von Regressionsmodellen im Kontext der multiplen Imputation stellen dabei nur zwei Beispiele relevanter Entwicklungen der Regressionsimputation dar. Damit lässt sich zwar die einfache Regressionsimputation eindeutig als traditionelles Verfahren definieren, allerdings stellen die Anpassungen durchaus fortgeschrittene Methoden dar.

Nearest-Neighbor-Methode

Die Nearest-Neighbor-Methode stellt ein weiteres modellbasiertes, aber auch grundlegendes Imputationsverfahren dar. In der Literatur wird dieses Verfahren zumeist als k-Nearest Neighbors (kNN) (MIT Critical Data 2016) oder allgemeiner als Imputation auf Basis von Distanzeigenschaften (Bankhofer 1995) bezeichnet. Zum Füllen der fehlenden Werte eines Objektes werden eine bestimmte Anzahl (k) nächster Nachbarobjekte des unvollständigen Objektes anhand der Distanzen zwischen den beobachteten Werte der Objekte bestimmt. Der Durchschnittswert der auszufüllenden Variable der k nächsten Nachbarn wird folglich als Wert für das unvollständige Objekt übernommen (MIT Critical Data 2016).

2.5.2 Moderne Imputationsmethoden

Nachdem im vorherigen Abschnitt die grundlegenden Imputationsverfahren erklärt wurden, werden im Folgenden die modernen bzw. fortgeschritteneren Verfahren vorgestellt. Namentlich werden die Multiple Imputation, Imputation auf Grundlage des Expectation-Maximization-Algorithmus und die Maximum-Likelihood-Imputation als State-of-the-Art-Methoden beschrieben. Dabei handelt es sich um diejenigen Methoden, die zwar in der Regel mit einem höheren Aufwand verbunden sind, gleichzeitig aber den Schwächen der traditionellen Methoden vorbeugen sollen.

Multiple Imputation

Die multiple Imputation wurde zwar bereits von Rubin erstmals im Jahr 1977 eingeführt, wurde aber seitdem stets weiterentwickelt und lässt sich daher durchaus als moderne bzw. fortgeschrittene Imputationsmethode einteilen. Die multiple Imputation zählt noch immer zu den State-of-the-art-Methoden (vgl. Baraldi & Enders 2010) und ist in einigen Domänen noch immer mitunter am weitesten verbreitet (vgl. Mistler & Enders 2017).

Die Grundidee der multiplen Imputation ist es, für jeden fehlenden Merkmalswert zwei oder mehr ($m \geq 2$) Imputationswerte zu generieren, um folglich auch $m \geq 2$ vervollständigte Datensätze zu erzeugen (Rubin 1987). Dabei existieren diverse Ansätze, um die Imputationswerte zu erzeugen (Donders et al. 2006). Daher handelt es sich bei der multiplen Imputation genau genommen nicht um eine einzelne, konkrete Methode, sondern um eine Sammlung von Techniken, die grundsätzlich demselben Vorgehen folgen. Dieses Vorgehen wird in der Regel in die Imputations-, die Analyse- und die Zusammenführungsphase unterteilt (Baraldi & Enders 2010; Lang & Little 2018). In der Imputationsphase werden für jeden fehlenden Wert zunächst m Werte mithilfe einer ausgewählten Imputationsmethode erzeugt und in den Datensatz ein-

gesetzt. In der Analysephase werden diese m vervollständigten Datensätze mithilfe der gewählten Methode analysiert, bevor die Imputationswerte bzw. die Datensätze in der Zusammenführungsphase aggregiert werden (Lang & Little 2018).

Mistler und Enders (2017) führen weiterhin aus, dass die Methoden der multiplen Imputation heutzutage häufig in die Joint Model (JM) Imputation und die Fully Conditional Specification (FCS) Imputation unterteilt werden. Nach dieser Unterscheidung beschreibt JM das simultane Einsetzen aller Fehlenden Werte, wohingegen FCS die aufeinanderfolgende Ergänzung der fehlenden Werte beschreibt. FCS wird häufig auch als Multiple Imputation by Chained Equations (MICE) bezeichnet (Grigorova et al. 2022). JM ist in der Literatur häufig auch als Multivariate Normal Imputation (MVNI) zu finden (vgl. Kalaycioglu et al 2016).

Durch die wiederholte Imputation und anschließende Analyse verspricht die multiple Imputation, den Nachteilen traditioneller Imputationsmethoden vorzubeugen. Konkret sorgt die richtige Anwendung der multiplen Imputation, anders als bei den vorgestellten traditionellen Methoden, dafür, dass wichtige statistische Eigenschaften, wie z. B. die Standardabweichung oder die Varianz, des ursprünglichen Datensatzes erhalten bleiben.

Maximum Likelihood Estimation und Expectation Maximization

Neben der multiplen Imputation gilt die Maximum Likelihood (ML) Estimation als zweite empfehlenswerte Imputationsmethode (Baraldi & Enders 2010). In der Literatur sind häufig auch Weiterentwicklungen dieser Methode als *Full Information Maximum Likelihood Estimation* (Little et al. 2014; Lang & Little 2018) oder *Maximum Likelihood Methods using the Expectation Maximization Algorithm* (vgl. Pigott 2001) zu finden. Laut Little et al. (2014) lassen sich mit diesen Methoden, ähnlich wie bei der multiplen Imputation, plausible Imputationswerte erzeugen, ohne dabei das statistische Verhalten des Datensatzes zu verzerren. Dabei ist im Gegensatz zur multiplen Imputation kein iteratives, mehrphasiges Vorgehen von Nöten. Außerdem bietet die ML-Imputation trotz der komplex erscheinenden mathematischen Beschreibung den Vorteil, dass Sie unter der Zuhilfenahme technischer Mittel ohne großen Aufwand umzusetzen ist (Baraldi & Enders 2010). Demzufolge verfolgt die ML-Imputation prinzipiell das Ziel, für fehlende Werte jeweils den zu erwartenden Wert bzw. den wahrscheinlichsten Wert auf Grundlage einer sogenannten log-likelihood-Funktion einzusetzen. Eine detailliertere Beschreibung dieses Verfahrens ist im Rahmen dieser Arbeit nicht weiter relevant.

Weitere modellbasierte Imputationsmethoden

Neben den gerade vorgestellten, modernen und modellbasierten Methoden zum Füllen fehlender Werte, existieren außerdem quasi unendlich viele weitere modellbasierte Imputationsverfahren. Die explizit vorgestellten Methoden beschreiben dabei nur die gängigsten der modellbasierten Methoden. Wie bereits zuvor ausgeführt, umfassen modellbasierte Imputationsmethoden diejenigen Verfahren, bei denen Vorhersagemodelle anhand der vorhandenen Werte entwickelt und anschließend für die Abschätzung der fehlenden Werte genutzt werden (vgl. MIT Critical Data 2016). Bei diesen Verfahren werden also Modellierungs- bzw. Data-Mining-Techniken innerhalb der Datenvorverarbeitung genutzt, ohne dabei direkt der Datenanalyse im Kontext der Wissensentdeckung in Datenbanken zu dienen. Da sich in der Datenwissenschaft neben den bereits vorgestellten Modellen (Nearest Neighbor, Regression) eine Vielzahl weiterer Vorhersagemodelle entwickelt haben und zudem ständig neue Modellierungsmöglichkeiten entwickelt werden, existieren dementsprechend auch genauso viele modellbasierte Imputationsverfahren. Weitere Nennenswerte und gängige Modellierungsmöglichkeiten zur Vorhersage von Werten, die an dieser Stelle nicht im Detail beschrieben werden, sind beispielsweise Entscheidungsbaummodelle und diverse weitere Machine-Learning-Implementationen.

3 Systematische Literaturrecherche zum Vergleich von Imputationsverfahren

Das zweite Kapitel dieser Arbeit hat den aktuellen Stand der Technik zu relevanten Bereichen der Datenwissenschaft und insbesondere zu Imputationsverfahren dargelegt. Dieses Kapitel behandelt nun mit dem Vergleich der zuvor präsentierten Imputationsverfahren. Damit adressiert dieses Kapitel explizit die Fragestellung dieser Arbeit nach der Vergleichbarkeit von Imputationsverfahren. Dazu wird eine systematische Analyse der bisherigen Literatur durchgeführt, deren Ziel es ist, Imputationsverfahren miteinander vergleichen zu können und zu bewerten. Die Methodik zur Auswertung der Literatur wird in Abschnitt 3.1 ausführlich beschrieben, bevor die eigentliche Auswertung in Abschnitt 3.2 vorgenommen wird.

3.1 Methodik der systematischen Literaturrecherche

Um Imputationsverfahren bewerten zu können, wird eine systematische bzw. strukturierte Literaturrecherche durchgeführt. Dabei handelt es sich um eine wissenschaftliche Methode, die ursprünglich aus dem Gesundheitswesen stammt, aber auch auf andere Forschungsbereiche übertragbar ist (vgl. Fink 2005). Finks (2005) Definition einer systematischen Literaturrecherche besagt, dass ebendiese mehrere Kriterien erfüllen muss: Sie muss einem systematischen Vorgehen unterliegen, dessen Ablauf genau beschrieben werden muss und darüber hinaus alle relevanten Veröffentlichungen umfasst. Diese Kriterien gewährleisten die Reproduzierbarkeit der Ergebnisse und erfüllen damit die Ansprüche an wissenschaftliches Arbeiten.

Das Vorgehen für diese Arbeit richtet sich nach einem Vorgehen von Okoli aus dem Jahr 2015. Okoli (2015) schlägt in seiner Veröffentlichung ein einheitliches Vorgehen für systematische Literaturrecherchen vor, indem er vorherige Methodiken darstellt, Elemente daraus übernimmt und zu einer standardisierten Methodik zusammenführt. Dieses Vorgehen bietet den Vorteil, dass es sich um einen modernen Ansatz handelt, der vorherige Ansätze allerdings nicht ausschließt, sondern deren Elemente und insbesondere deren Vorzüge zusammenfasst. Darüber hinaus hält Okoli seine Richtlinien so allgemein wie möglich, damit sie auf diverse Forschungsgebiete übertragen werden können. Zudem erlaubt das Vorgehen im Gegensatz zu anderen Ansätzen den Einbezug von sowohl qualitativer als auch quantitativer Studien.

Wie mithilfe von Abbildung 3-1 veranschaulicht, umfasst die systematische Literaturrecherche nach Okoli vier übergeordnete Phasen, die sich weitergehend in acht Schritte unterteilen lassen. Die einzelnen Phasen und Schritte werden an dieser Stelle kurz erläutert, bevor sie, sofern möglich, in den Abschnitten 3.1.1 bis 3.1.3 konkret auf den Untersuchungsgegenstand dieser Arbeit übertragen werden.

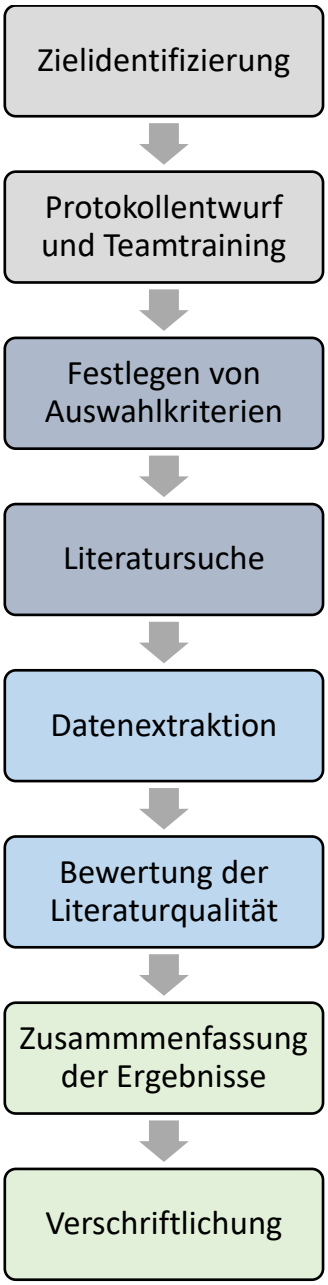

Abbildung 3-1: Nach Phasen unterteilter Ablauf der systematischen Literaturrecherche (eigene Darstellung in Anlehnung an Okoli (2015, S. 885))

Die Planungsphase beinhaltet die Zielidentifizierung (1) und das Erstellen eines Protokollentwurfs sowie das Vorbereiten der beteiligten Personen (2). Die Auswahlphase beinhaltet das erstmalige Festlegen von Auswahlkriterien (3) und folglich die eigentliche Suche bzw. Auswahl relevanter Literatur (4). Die dritte Phase umfasst die Gewinnung von neuen Erkenntnissen durch das Extrahieren von Daten (5) und dem Bewerten der Literaturqualität (6). Die Durchführung umfasst schließlich das Zusammenfassen der Ergebnisse (7) und das Ausformulieren der gesamten Literaturanalyse (8).

3.1.1 Planung der systematischen Literaturrecherche

Die Identifizierung und Dokumentation der Ziele stellt, wie zuvor beschrieben, den ersten Schritt der Planungsphase dar. Die Ziele dieser Literaturrecherche sind eng mit den Zielen der Arbeit verknüpft. Konkret adressiert die systematische Literaturrecherche in diesem Kapitel die Fragestellung nach der Vergleichbarkeit von Imputationsverfahren, die sich wiederum in zwei untergeordnete Teilziele aufteilen lässt. Einerseits verfolgt die systematische Literaturrecherche das Ziel, abbilden zu können, wie sich Imputationsverfahren miteinander vergleichen lassen. Andererseits wird das Ziel verfolgt, einzelne Verfahren oder Verfahrensgruppen qualitativ bewerten zu können. Es wird also konkret nach der Eignung von Imputationsverfahren, auch im Kontext von Wissensentdeckungsprozessen in Datenbanken, gefragt. Die hier vorgenommene Literaturrecherche hat dabei den Anspruch, den Forschungsgegenstand eigenständig zu beleuchten und neue Erkenntnisse zu generieren, ohne diese anschließend anwenden zu müssen.

Der ebenfalls in der Planungsphase enthaltene Protokollentwurf und die Vorbereitung der an der Recherche beteiligten Personen zielt zwar nach Okoli (2015) insbesondere auf die strukturierte Zusammenarbeit in Gruppen ab, allerdings ist die Protokollierung auch für selbständiges Arbeiten durchaus entscheidend. Eine einheitliche Protokollierung stellt demzufolge das systematische Vorgehen sicher, indem es Autoren davon abhält, in alternative Vorgehensweisen zu verfallen. Die Fortschritte der hier vorgenommenen Literaturrecherche werden daher in einem Excel-Arbeitsblatt festgehalten, dass nach Sichtung der Literatur auch der Aggregation der Ergebnisse dienen soll. In diesem Dokument werden also neben grundsätzlichen Informationen wie dem Titel, dem Autor und dem Erscheinungsjahr auch relevante Inhalte der Veröffentlichungen festgehalten.

3.1.2 Literaturauswahl

Die Festlegung und Dokumentation der Auswahlkriterien ist ein entscheidender Schritt einer systematischen Literaturrecherche, da dadurch die eigentliche Durchführbarkeit und darüber hinaus die Reproduzierbarkeit der Ergebnisse gewährleistet wird. Hier geht es nicht darum, die Qualität der Veröffentlichungen wertend zu beurteilen, sondern pragmatische Kriterien für eine Vorauswahl festzulegen. Okoli (2015) benennt unter anderem den Inhalt, die Sprache, die Zugriffsmöglichkeit, die Art der Veröffentlichung, bestimmte Autoren, das Setting der Studien, das Studiendesign und das Veröffentlichungsdatum als Kriterien für die Vorauswahl. Darüber hinaus gibt es weitere Kriterien, die für diese Arbeit allerdings keine Relevanz haben.

Der gesuchte Inhalt und die damit verbundenen Suchwörter sind stark von den Zielen der Literaturrecherche abhängig. Da diese Arbeit keine konkreten, sondern alle Imputationsmethoden zum Zweck der Datenvorverarbeitung betrachtet und nach der Vergleichbarkeit der Methoden fragt, wird in erster Instanz nach allen Veröffentlichungen gesucht, in der Imputationsmethoden miteinander verglichen werden. Da die rein formalen Vor- und Nachteile einzelner Verfahren und Verfahrensgruppen schon in Abschnitt 2.5 thematisiert wurden, wird die systematische Literaturrecherche auf praktische Vergleichsstudien begrenzt. Um ausschließlich Veröffentlichungen zu finden, deren Fokus auf der Durchführung einer Vergleichsstudie liegt, wird nach Arbeiten gesucht, deren Titel bestimmte Stichwörter bzw. Stichwortkombinationen enthält. Konkret wird nach Veröffentlichungen gesucht, deren Titel eine Kombination der Stichwörter *Vergleich, Vergleichsstudie, Studie, Imputation, Imputationsmethoden* und *Imputationstechniken* bzw. der englischsprachigen Entsprechungen enthält. Um dabei gleichzeitig veraltete Ergebnisse auszuschließen und die Suche auf eine begrenzte Anzahl an Veröffentlichungen einzugrenzen, werden im ersten Schritt nur Veröffentlichungen der letzten zehn Jahre eingeschlossen. In einem weiteren Schritt werden dann in Jahresschritten weiter zurückliegende Ergebnisse betrachtet, um festzustellen, ob daraus neue Erkenntnisse gewonnen werden können. Sofern diese Suchergebnisse keine neuen Erkenntnisse liefern, wird die Suche und Auswertung an dieser Stelle beendet. Da sich gemäß dem zweiten Kapitel ausschließlich alle Verfahren auf numerische Datensätze anwenden lassen und Datensätze mit nominalen Daten in diesem Kontext eine absolute Seltenheit sind, wird darüber hinaus ausdrücklich nach Studien gesucht, die ausschließlich numerische Datensätze verwenden. Dieses Kriterium ermöglicht im Optimalfall eine quantitative Aggregation der Ergebnisse. Um keine hilfreichen Veröffentlichungen auszuschließen, werden per se weder bestimmte Autoren noch Domänen noch bestimmte Arten von Veröffentlichungen ausgeschlossen. Auch beim Studiendesign wird keine Vorauswahl getroffen, da es laut der Zielsetzung auch darum geht, verschiedene Vorgehensweisen zum Vergleich von Imputationsverfahren herauszuarbeiten. Um eine möglichst große Anzahl potenziell hilfreicher Veröffentlichungen zu betrachten, wird Google Scholar als Datenbank für die Literatursuche genutzt. Google Scholar bietet den Vorteil, dass diverse, online verfügbare Datenbanken nach Literatur durchsucht werden und so die Veröffentlichungen verschiedener Quellen für wissenschaftliche Literatur gebündelt werden. Aus Gründen der Verständlichkeit und Zugänglichkeit werden nur Veröffentlichungen in deutscher und englischer Sprache herangezogen, die entweder frei oder über den institutionellen Zugang der Universitätsbibliothek Dortmund zugänglich sind.

3.1.3 Qualitative Selektion der Literatur und Aggregation der Ergebnisse

Nachdem die relevanten Veröffentlichungen mithilfe der in Abschnitt 3.1.2 vorgestellten, formalen Kriterien vorselektiert werden, werden die Arbeiten näher untersucht und ausgewertet, um den Inhalt und die Ergebnisse der relevanten Arbeiten schließlich zu aggregieren. In einem ersten Schritt werden die einzelnen Veröffentlichungen dazu anhand mehrerer, qualitativer Kriterien auf die Eignung für diese Arbeit geprüft, die an dieser Stelle kurz erläutert werden.

Bei einer ersten inhaltlichen Sichtung der Veröffentlichungen wird geprüft, ob es sich überhaupt um Arbeiten zum Thema der Imputation von fehlenden Merkmalswerten handelt. Dazu wird, sofern vorhanden, die Kurzfassung bzw. das Abstract auf die Anforderungen dieser systematischen Literaturrecherche geprüft. Ist kein Abstract vorhanden, wird weitergehend die Einleitung und gegebenenfalls die Methodik der Studie auf relevante Informationen untersucht. Es bedarf dieser Überprüfung trotz der definierten Suchwörter, da der Imputationsbegriff selten auch in anderem Kontext, z. B. in der Biologie im Bereich der Genomimputation, verwendet wird.

In einem weiteren Schritt wird geprüft, ob in der Arbeit tatsächlich eine vergleichende Studie verschiedener Imputationsverfahren präsentiert wird. Dabei werden weitergehend nur diejenigen Studien berücksichtigt, deren Methodik explizit beschrieben wird und die anhand von definierten Auswertungsmetriken evaluiert werden. Außerdem unterliegen die verwendeten Studien dem Anspruch, dass die Imputation der Datenvorverarbeitung dienen. Insbesondere bei der Verwendung von longitudinalen Datensätzen in Kombination mit modellbasierten Imputationsmethoden ist eine Abgrenzung zwischen Data Mining und Imputation häufig schwierig, da die Imputation in diesen Fällen direkt der Wissensgewinnung dient. Daher werden nur diejenigen Studien mit longitudinalen Datensätzen einbezogen, bei denen keine speziellen Algorithmen für longitudinale Daten angewendet werden und bei denen die Daten eindeutig zur Datenvorverarbeitung vervollständigt werden. Um die Ergebnisse der Studien aggregieren und übertragen zu können, werden außerdem nur diejenigen Studien verwendet, in denen nicht ausschließlich domänen- oder anwendungsspezifischen Verfahren benutzt werden.

3.2 Synthese der Vergleichsstudien

In diesem Kapitel werden die Ergebnisse der systematischen Literaturrecherche beschrieben und ausgewertet, indem die hauptsächlichen Fragestellungen zur Vergleichbarkeit von Imputationsverfahren getrennt voneinander referenziert werden. In Abschnitt 3.2.1 wird das Vorgehen der Vergleichsstudien miteinander verglichen, indem die Studiendesigns, die in den Studien untersuchten Faktoren und die Auswertungsmetriken herausgearbeitet werden. Dieser Abschnitt bezieht sich damit direkt auf die Fragestellung, wie Imputationsverfahren miteinander verglichen werden können. Das ultimative Ziel dieses Abschnittes ist es, ein allgemein anwendbares Rahmenkonzept zum Vergleich von Imputationsverfahren anhand der betrachteten Studien herauszuarbeiten. In Abschnitt 3.2.2 werden dann die eigentlichen Ergebnisse der Vergleichsstudien zusammengefasst, indem die Leistungsfähigkeit der Verfahren anhand der zuvor dargestellten Auswertungsmetriken bewertet wird. Dieser Abschnitt referenziert damit die Frage nach der Qualität der verschiedenen Imputationsverfahren.

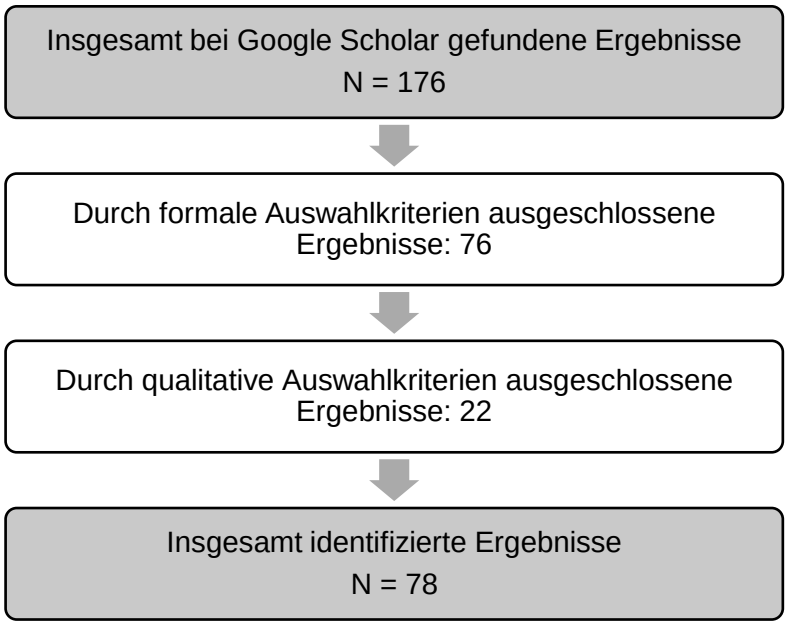

Abbildung 3-2: Literaturauswahlprozess

Mithilfe der in 3.1.2 und 3.1.3 definierten formalen und qualitativen Auswahlkriterien konnten, wie in Abbildung 3-2 veranschaulicht, insgesamt 78 Veröffentlichungen identifiziert werden, deren Inhalt alle gestellten Ansprüche erfüllt. Wie in Abschnitt 3.1.2 erläutert, wurde außerdem erwägt, den Betrachtungszeitraum schrittweise auszuweiten, um gegebenenfalls weitere Ergebnisse zu erhalten. Nach Durchsicht weiterer Ergebnisse konnte allerdings schnell festgestellt werden, dass ältere Veröffentlichungen keinerlei Mehrwert in Form von alternativen Studiendesigns, untersuchten Faktoren oder Auswertungsmetriken darbieten konnten. Daher werden im Folgenden die Inhalte der 78 identifizierten Studien aus den Jahren 2013 bis 2023

aggregiert. Dabei ist jedoch anzumerken, dass wegen des unterschiedlichen Aufbaus nicht alle Studien in die Aggregation aller betrachteten Aspekte eingehen können. Zur übersichtlichen Darstellung und zur einfachen Auswertung wurden die untersuchten Inhalte der Veröffentlichungen in eine Literaturtabelle übertragen. Die Tabelle wurde zur besseren Darstellung für diese Arbeit in zwei Tabellen aufgeteilt, die an dieser Stelle auszugsweise und im Anhang vollständig abgebildet sind. Die auszugsweise Darstellung dient dem einfacheren Verständnis dafür, anhand welcher Kriterien die Ergebnisse für diese Arbeit zusammengefasst werden. Dabei stellt Tabelle 1 eine Art Verzeichnis für Tabelle 2 dar, das Titel, Autor, Erscheinungsjahr und allgemeine Domäne der jeweiligen Veröffentlichungen enthält. Tabelle 2 dient weitergehend als Auswertungstabelle für die Synthese der Studiencharakteristika. Diese Tabelle enthält daher Spalten der relevanten Charakteristika, die in Abschnitt 3.2.1 aggregiert werden. Die Ergebnisse der Studien, die Aufschluss über die Leistungsfähigkeit und Qualität der Imputationsverfahren geben, werden anschließend anhand der Charakteristika aus Abschnitt 3.2.1 differenziert zusammengefasst. Diese Zusammenfassung kann bestenfalls als Grundlage für Handlungsempfehlungen für die Auswahl eines geeigneten Imputationsverfahrens für bestimmte Anwendungsfälle dienen.

Tabelle 1: Auszug aus der Literaturtabelle der systematischen Literaturrecherche

	Titel	Autor	Jahr	Domäne
[1]	Comparison of imputation methods for missing laboratory data in medicine	Waljee et al.	2013	Medizin
[2]	Missing traffic data: comparison of imputation methods	Li, Li & Li	2014	Verkehr
[3]	Comparison of performance of data imputation methods for numeric dataset	Jadhav, Pramod & Ramanathan	2019	Datenwissenschafft
[4]	A comparison of various imputation methods for missing values in air quality data	Zainuri, Jemain & Muda	2015	Energie und Umwelt
[5]	Missing network data a comparison of different imputation methods	Krause et al.	2018	Energie und Umwelt
[6]	Comparison of missing value imputation methods in time series: the case of Turkish meteorological data	Yozgatligil et al.	2011	Energie und Umwelt
[7]	Comparison of imputation methods for missing values in air pollution data: Case study on Sydney air quality index	Wijesekara & Liyanage	2020	Energie und Umwelt
[8]	Comparison of five iterative imputation methods for multivariate classification	Liu & Brown	2012	Datenwissenschafft
[9]	Comparison of imputation methods for end-user demands in water distribution systems	Jun, Jung & Lansey	2021	Energie und Umwelt
[10]	A comparison of multiple imputation methods for missing data in longitudinal studies	Huque et al.	2018	Gesundheit und Soziales

Tabelle 2: Auszug aus der Auswertungstabelle der systematischen Literaturrecherche

		Typ		Datenursprung			Vollständigkeit				Ausfallmechanismus					Ausfallrate			Auswertungsmetriken		
	Veröffentlichung	longitudinal	nicht longitudinal	Anzahl verwendeter Datensätze	real	simuliert	real und simuliert	vollständig	unvollständig	verwendung der vollständigen Objekte	vollständige und unvollständige	nicht explizit angegeben	MCAR	MAR	MNAR	gemischt	Aufallmechan. nicht explizit angegeben	Ausfallmechan. als untersuchter Faktor?	verwendete Ausfallraten in Prozent	Ausfallrate nicht explizit angegeben	Ausfallrate als untersuchter Faktor?	statistisches Verhalten	Vorhersagegenauigkeit	Modellierungsfähigkeit	weitere Bewertungskriterien
[1]			x	2	x			x									x		10, 20, 30		x		x	x	
[2]		x		1	x			x					x	x		x			5 bis 50		x		x		
[3]			x	6	x			x									x		10, 20, 30, 40, 50		x		x		
[4]		x		1	x			x									x		5, 10, 15, 20, 25, 30		x		x		
[5]			x	1		x		x					x	x	x	x		x	10, 20, 30, 40, 50		x	x			
[6]		x		1	x					x							x		10, 20, 50		x		x		
[7]		x		1	x				x				x						5, 10, 15, 20		x	x			
[8]			x	3			x	x					x						10 (s); 5, 10, 15, 20, 25, 30, 35, 40 (r)		x	x		x	
[9]		x		1	x			x		x							x		0,25; 0,5; 10; 20; 30; 40; 50		x			x	
[10]		x		2			x	x			x			x						x		x	x		
[11]		x		2	x			x			x						x		5, 10, 15, 20, 25, 30		x		x	x	
[12]			x	1	x			x					x						ca. 20			x			
[13]			x	3	x			x	x					x					3,85; 11,95; 23,52		x		x	x	
[14]		x		14			x	x					x						5, 10, 20, 30, 40, 50		x			x	
[15]		x		1	x		x	x									x	x	20, 50		x		x		

3.2.1 Vorgehensweisen zum Vergleich von Imputationsverfahren

Dieser Abschnitt fokussiert die Art und Weise, mit denen Imputationsverfahren miteinander verglichen werden können. Dazu werden zunächst die in den untersuchten Studien verwendeten Methodiken qualitativ zusammengefasst und verglichen. Dabei werden auch die Rahmenbedingungen bzw. der Kontext der Studien kurz beleuchtet. Anschließend werden die in den Studien untersuchten Faktoren herausgearbeitet. In einem weiteren und für diese Arbeit sehr entscheidenden Schritt werden die Auswertungsmetriken der Studien erörtert. In diesem Zuge werden die Metriken benannt, nach Kategorien unterteilt und übersichtlich eingeordnet.

Studiendesigns

Die Studiendesigns zur Untersuchung und zum Vergleich von Imputationsverfahren werden in der Regel durch die Methodik der Vergleichsstudien abgebildet. Da das jeweilige Studiendesign von einer Vielzahl von Rahmenbedingungen und Einflussfaktoren abhängt, gleicht kein Studiendesign dem anderen. Daher stellt die qualitative Synthese der Methodiken zum Vergleich von Imputationsverfahren in diesem Abschnitt nur eine Möglichkeit dar, die Studiendesigns zusammenzufassen. Entscheidende Rahmenbedingungen bzw. Einflussfaktoren können z. B. die Zielsetzung der Studie, der Datenursprung, die Anzahl der verwendeten Datensätze, die Merkmale fehlender Daten und die untersuchten Faktoren sein. Im Folgenden werden diese Rahmenbedingungen und einflussgebenden Faktoren der identifizierten Studien mitsamt deren Konsequenzen für die Entwicklung einer Studie kurz erklärt, bevor das Vorgehen für Vergleichsstudien über Imputationsverfahren übersichtlich zusammengefasst wird. Dabei ist vorab anzumerken, dass eine Unterscheidung zwischen Einfluss- und Untersuchungsfaktor nicht immer ganz einfach. Einzelne Faktoren können in der einen Studie als Rahmenbedingung dienen und in anderen Studien expliziter Untersuchungsgegenstand sein. Daher werden einzelne Faktoren sowohl hier als Rahmenbedingung als auch später als untersuchter Faktor aufgeführt.

So stellen die Merkmale fehlender Daten, also das Ausfallmuster, die Ausfallrate und der Ausfallmechanismus in einigen Studien eine reine Rahmenbedingung dar, die z. B. als Begründung für die Auswahl der untersuchten Imputationsverfahren herangezogen wird. Andererseits können diese Merkmale und deren Einfluss auf das Ergebnis der Imputation auch explizit untersucht werden, indem vollständige Datensätze so manipuliert werden, dass sie verschiedene Ausfallmuster, -raten oder -mechanismen abbilden.

Bezüglich des Forschungsziels lassen sich die untersuchten Vergleichsstudien grundsätzlich nach zwei übergeordneten Zielen unterscheiden. Einerseits fokussieren einige Studien explizit Imputationsverfahren und verfolgen diesbezüglich das Ziel, Imputationsverfahren möglichst objektiv zu bewerten und damit domänenübergreifende Handlungsempfehlungen abzuleiten.

Diese Studien, die insgesamt 20 der 78 identifizierten Veröffentlichungen ausmachen, leisten damit direkt einen Beitrag für den Bereich der Datenwissenschaft. Auf der anderen Seite werden Imputationsverfahren in den verbleibenden drei Vierteln der Studien als Werkzeug genutzt, um domänenspezifische Daten zu vervollständigen. Die in den identifizierten Veröffentlichungen am häufigsten betrachtete Domäne ist die *Medizin*, aus der die Datensätze für 22 Studien stammen, wodurch die Medizin allein einen größeren Anteil ausmacht als die Datenwissenschaft. Datensätze aus dem Bereich *Energie und Umwelt* werden in insgesamt 17 Veröffentlichungen und damit nur knapp seltener verwendet. Elf Datensätze stammen aus dem Feld *Gesundheit und Soziales*. Darunter sind diejenigen Studien zusammengefasst, die meist durch Umfragen erzeugte und personenbezogene Datensätze verwenden. Die verbleibenden acht Veröffentlichungen entstammen verschiedenen Domänen und Wissenschaftsfeldern. Dabei ist auffällig, dass nur drei der 78 Studien aus dem industriellen Kontext und gar keine der Studien aus dem wirtschaftswissenschaftlichen Bereich stammen. Denn wie im zweiten Kapitel dargelegt existieren für diese Domänen sogar spezielle Vorgehensmodelle für die Wissensentdeckung in Datenbanken, die jedoch in keiner der betrachteten Studien referenziert werden. All die Studien, die nicht direkt im Kontext der Datenwissenschaften verfasst wurden, liefern nicht unbedingt übertragbare Ergebnisse und verfolgen vordergründig das Ziel, Imputationsverfahren im Kontext der eigenen Domäne oder gar nur für den konkreten Anwendungsfall zu bewerten. Bei Studien mit der zweitgenannten Zielsetzung ist einzeln zu prüfen, inwiefern sich aus den jeweiligen Ergebnissen objektive Handlungsempfehlungen ableiten und auf andere Domänen übertragen lassen.

Ein Merkmal, das sich bei ausnahmslos allen Studien untersuchen lässt und Aufschluss über das Vorgehen gibt, ist die Anzahl der Datensätze, die in der jeweiligen Studie zur Bewertung der Verfahren genutzt werden. Die Verwendung mehrerer Datensätze wird in denjenigen Studien, die mehr als einen Datensatz verwenden, dadurch begründet, dass sie durch einen größeren Stichprobenumfang valide und folglich repräsentative Ergebnisse erzeugen. Trotz dieser Argumentation bauen mehr als zwei Drittel der Studien und deren Ergebnisse auf nur einem einzigen Datensatz auf. Dabei ist anzumerken, dass die verwendete Datenbasis in sechs dieser 55 Studien aus mehreren Datensätzen zusammengesetzt ist. Sieben Studien nutzen immerhin zwei und 16 Studien nutzen drei oder mehr Datensätze zur Bewertung der Verfahren. Ferner stechen zwei Studien dadurch heraus, dass zehn oder mehr Datensätze zum Vergleich von Imputationsverfahren vervollständigt werden.

Der Datenursprung wird in nahezu allen Studien thematisiert und lässt sich auf zwei Ebenen unterscheiden. Zunächst kann zwischen simulierten und realen Daten unterschieden werden. Weitergehend lassen sich Realdaten nach dem tatsächlichen Ursprung der Daten bzw. nach

der gerade schon angesprochenen Domäne, aus der die Daten stammen, unterteilen. Während die Unterscheidung zwischen realen und simulierten Daten einen direkten Einfluss auf den jeweiligen Ablauf der Studien hat, ist die Domäne eher als Rahmenbedingung anzusehen, die sich gegebenenfalls auf die Auswahl der zu untersuchenden Faktoren und die Zielsetzung der Studien auswirkt. Insgesamt bauen 58 der 78 betrachteten Studien ausschließlich auf Datensätzen aus der Realität auf. Neun Studien verwenden ausschließlich simulierte Datensätze und weitere elf Studien untersuchen sowohl simulierte als auch reale Datensätze. Der Unterschied zwischen realen und simulierten Daten beeinflusst, ob der Datensatz für die Studien per se vollständig oder unvollständig ist, was sich wiederum auf das Vorgehen für die jeweilige Studie auswirkt. Denn je nachdem, ob der ursprüngliche Datensatz vollständig ist oder nicht, bedarf es einer dementsprechenden Anpassung des Studiendesigns. Sofern der ursprüngliche Datensatz unvollständig ist, was in der Regel bei realen Datensätzen auftritt, existieren zwei Optionen für das weitere Vorgehen zum nachfolgenden Vergleich von Imputationsverfahren. Einerseits kann der unvollständige Datensatz direkt mithilfe von Imputationsverfahren vervollständigt werden, was jedoch den Nachteil mit sich bringt, dass die Auswertung der Studie erschwert wird. Wegen dieser Nachteile im Bereich der Auswertung werden Imputationsverfahren in nur elf Studien anhand von vorneherein unvollständigen Datensätzen untersucht. Andererseits verfolgen knapp 80 der Studien die Strategie, im Vorfeld der Imputation Werte zu entfernen und anschließend mithilfe von Imputationsverfahren wieder auszufüllen. In elf der Studien 62 Studien mit vollständigen Datensätzen ist weitergehend angegeben, dass die verwendeten Datensätze zwar ebenfalls fehlende Werte aufweisen, dass die final untersuchten Datensätze allerdings nur aus den vollständig beobachteten Objekten gebildet wurden. Damit besteht bei der Verwendung realer, unvollständiger Daten die Option, nur die vollständigen Objekte für die weitere Studie zu verwenden. Dieses Vorgehen birgt zwar Vorteile für die Vergleichbarkeit, stellt jedoch gleichzeitig auch Probleme für die Studie dar. Denn wie im zweiten Kapitel dargelegt, können bei der Eliminierung ganzer Objekte zum einen wichtige Informationen der Daten verloren werden und zum anderen kann dieses Vorgehen in der Regel nicht für longitudinale Daten angewendet werden, da dadurch Lücken in den zeitlich abhängigen Datensätzen entstünden.

Die gerade angesprochene Unterscheidung zwischen longitudinalen und nicht-longitudinalen Daten hat, wie bereits im zweiten Kapitel ausgeführt, in erster Linie Auswirkungen auf die ausgewählten Imputationsverfahren und die damit verbundenen Auswertungsmetriken. Während 43 der Studien nicht-longitudinale Daten zur Untersuchung von Imputationsverfahren nutzen, werden 34 Studien longitudinale Daten zugrunde gelegt. Lediglich eine Studie verwendet sowohl longitudinale als auch nicht-longitudinale Datensätze.

Der Kern jeder hier betrachteten Studie, der letzten Endes dem Vergleich der Imputationsverfahren dient, enthält allein durch die Auswahlkriterien für die systematische Literaturrecherche immer dieselben Elemente. Grundsätzlich beginnt der eigentliche Vergleich mit einem unvollständigen Datensatz, auf den folglich verschiedene Imputationsverfahren angewendet werden, wodurch ein vervollständigter bzw. imputierter Datensatz erzeugt wird. Einige der untersuchten Studien gehen darüber hinaus einen Schritt weiter und verwenden die imputierten Datensätze im Folgenden zur Modellierung bzw. zum Data Mining mit dem Ziel der Wissensgewinnung. Insgesamt 19 von 78 und damit rund ein Viertel betrachteten Studien verwenden die imputierten Datensätze zum anschließenden Data Mining und stehen damit implizit im Kontext der Wissensentdeckung in Datenbanken. Unabhängig vom Schritt der Modellierung werden die Ergebnisse der Studie anschließend in irgendeiner Form ausgewertet und miteinander verglichen. Die Auswertungskriterien hängen dabei wiederum von den zuvor beschriebenen Rahmenbedingungen, den untersuchten Faktoren und dem Studiendesign selbst ab. Die in den Studien verwendeten Metriken werden in einem nachfolgenden Abschnitt gesondert beleuchtet und später in das vorgeschlagene Rahmenkonzept integriert.

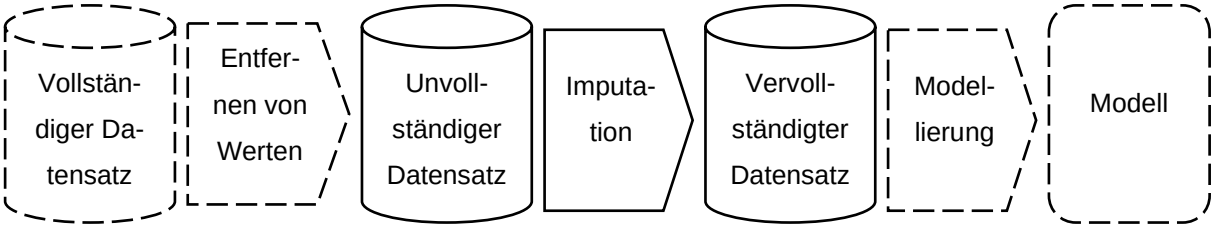

Abbildung 3-3: Zusammenfassung der Studiendesigns zum Vergleich von Imputationsverfahren

Abbildung 3-3 verdeutlicht den gerade dargelegten, grundsätzlichen Ablauf aller Studien samt der optionalen Ablaufschritte. Dabei sind die optionalen Ablaufschritte, die nicht zwangsläufig notwendig sind, gestrichelt abgebildet, während die Kernelemente einer jeden Studie mit durchgängigen Linien konturiert sind. Die Zylinder und das Rechteck repräsentieren in dieser Grafik die Zwischen- und Endergebnisse, die mithilfe der als Blockpfeile abgebildeten Ablaufschritte erreicht werden.

Untersuchte Faktoren

Wie im vorherigen Abschnitt dargelegt, lassen sich Imputationsverfahren mithilfe verschiedener, wenn auch ähnlicher Methodiken miteinander vergleichen. Dieser Abschnitt fokussiert nun die in den Studien untersuchten Faktoren auf den jeweils angestellten Vergleich. Denn neben dem Vergleich verschiedener Imputationsverfahren werden in nahezu allen betrachteten Studien weitere Parameter verändert, um den Einfluss weiterer Faktoren auf den Prozess zu untersuchen. Als untersuchte Faktoren werden in dieser Arbeit also diejenigen Faktoren verstanden, die bei der wiederholten Durchführung der Studien verändert werden. Dadurch werden in

jeder Studie diverse Faktorkombinationen aus den untersuchten Imputationsverfahren und den veränderlichen Untersuchungsfaktoren untersucht. Dabei stehen häufig auch die Wechselwirkungen zwischen ebendiesen Faktoren und den Imputationsverfahren im Vordergrund. So können Studien durch die Veränderung weiterer Parameter Aufschluss über die Eignung bestimmter Imputationsverfahren in Abhängigkeit des jeweiligen Anwendungsfalls geben. Eine Analyse dieser Faktoren ist im Kontext dieser Arbeit relevant, da diese zum Ableiten von Handlungsempfehlungen entscheidend sind. Im Folgenden werden daher, ähnlich wie zuvor die Rahmenbedingungen und Einflussfaktoren, die in den Studien untersuchten Faktoren herausgearbeitet.

Wie bereits im vorherigen Abschnitt zu den Studiendesigns dargelegt, können die Merkmale fehlender Daten, namentlich das Ausfallmuster, die Ausfallrate und der Ausfallmechanismus je nach Studie entweder eine reine Rahmenbedingung oder ein untersuchter Faktor sein. Da zu imputierende Datensätze in der Regel anhand dieser Merkmale beschrieben werden, werden diese Merkmale oder zumindest einige dieser Merkmale zur Beschreibung fast aller betrachteten Studien verwendet.

Während der Ausfallmechanismus und die Ausfallrate in nahezu jeder Studie zumindest als Rahmenbedingung und häufig auch als Untersuchungsfaktor thematisiert werden, ist das Ausfallmuster ein vergleichsweise seltener Untersuchungsgegenstand. Ein Vergleich verschiedener Ausfallmuster ist dabei vornehmlich im Kontext derjenigen Studien mit longitudinalen Daten von Bedeutung, wobei auch nur einzelne dieser Studien einen tatsächlichen Vergleich verschiedener Ausfallmuster anbieten. Bei longitudinalen Daten wird in den betrachteten Studien zumeist danach unterschieden, ob Werte ab einem gewissen Zeitpunkt gänzlich fehlen oder ob fehlende Werte zufällig und zeitlich unabhängig auftreten. Ohne dadurch genaue Aussagen über die Leistungsfähigkeit verschiedener Imputationsverfahren in Abhängigkeit von Ausfallmustern treffen zu können, lässt sich immerhin festhalten, dass Ausfallmuster und Ausfallmechanismus insbesondere bei longitudinalen Daten eng miteinander verknüpft sind.

Der Ausfallmechanismus wird im Gegensatz zum Ausfallmuster deutlich häufiger untersucht oder zumindest thematisiert. Konkret benennen 55 der 78 Studien den Ausfallmechanismus, wobei in elf dieser Studien ein Vergleich von Ausfallmechanismen bzw. deren Einfluss auf das Ergebnis der Imputation durchgeführt wird. In diesen Studien werden also entweder Werte nach mindestens zwei verschiedenen Ausfallmechanismen aus demselben vollständigen Datensatz entfernt, anschließend imputiert und miteinander verglichen oder aber es werden zwei verschiedene Datensätze mit verschiedenen Ausfallmechanismen gegenübergestellt. Weitere 44 Studien führen den Ausfallmechanismus zumindest als Rahmenbedingung auf bzw. unter-

suchen die Eignung verschiedener Imputationsverfahren für die gewählten Ausfallmechanismen. Unter den 55 Studien, die den Ausfallmechanismus angeben, sind MCAR- und MAR-Datensätze mit 33 respektive 30 Verwendungen fast gleichermaßen vertreten. Nur neun Studien testen Imputationsverfahren an MNAR-Datensätzen. Weitere drei Veröffentlichungen charakterisieren die verwendeten Datensätze durch Mischformen der Ausfallmechanismen. Dies kann beispielsweise auftreten, wenn fehlende Merkmalsausprägungen eines multivariaten Datensatzes verschiedenen Ursachen unterliegen oder wenn ein Datensatz aus verschiedenen Datensätzen zusammengesetzt wird. Mit 23 von 78 Studien benennen knapp ein Drittel der Studien den Ausfallmechanismus gar nicht. Dabei sind auch diejenigen Studien inbegriffen, in denen beispielsweise angeben wird, dass Werte zufällig entfernt werden, ohne jedoch einen konkreten Ausfallmechanismus anzugeben. Wird der Ausfallmechanismus im Kontext einer Studie weder verändert noch benannt, so ist dies insofern problematisch, dass die Ergebnisse der Studie weniger gut übertragbar sind.

Die Ausfallrate wird in den identifizierten Studien weitaus häufiger als der Ausfallmechanismus untersucht. Besonders in den Studien, in denen mit einem vollständigen Datensatz gestartet wird und aus dem im Vorfeld Imputation Werte entfernt werden, werden in der Regel verschiedene Ausfallraten und deren Einfluss auf die Imputationsergebnisse getestet. Insgesamt wird die Ausfallrate in rund 80% der Studien angegeben und in wiederum 80% dieser Studien als Untersuchungsfaktor verwendet. Innerhalb der Studien mit vollständigen Datensätzen wird der Einfluss der Ausfallrate sogar in 100% der Fälle untersucht. Im Umkehrschluss wird die Ausfallrate in nur drei aus elf Studien mit unvollständigen Datensätzen untersucht, in weiteren vier dieser Studien immerhin benannt und in den verbleibenden vier Studien gar nicht benannt. Ferner ist bei der Ausfallrate anzumerken, dass in einigen Veröffentlichungen lediglich eine Ausfallrate je Merkmal und damit keine oder nur eine ungefähre Ausfallrate bezogen auf den ganzen Datensatz angegeben wird. Diese Tatsache ergibt zwar einerseits Probleme beim Vergleich unter dem Einfluss verschiedener Ausfallraten, kann andererseits jedoch auch Anhaltspunkte zum Ausfallmuster geben, sofern dieses nicht explizit thematisiert wird.

Abseits der gerade aufgeführten Einfluss- und Untersuchungsfaktoren werden in einigen wenigen Studien weitere Faktoren untersucht. So wird zum Beispiel insbesondere bei simulierten Datensätzen der Einfluss der Korrelation zwischen den Variablen innerhalb eines Datensatzes untersucht werden, da es sich bei der Korrelation um einen möglichen Parameter bei der künstlichen Erzeugung von Datensätzen handelt. Von einer tieferen Zusammenfassung weiterer Untersuchungsfaktoren wird an dieser Stelle jedoch abgesehen, da die zu geringe Anzahl an Studien mit ebenjenen Untersuchungsfaktoren nicht zum Vergleich der Verfahren beitragen würde.

Auswertungsmetriken

In diesem Abschnitt werden nun die von den zuvor charakterisierten Studien verwendeten Auswertungsmetriken zusammengefasst. Die Betrachtung von Auswertungsmetriken ist für diese Arbeit aus mehreren Gründen von Bedeutung. Zunächst referenzieren Auswertungsmetriken das zuvor aufgeführte fünfte Ziel von Imputationsverfahren, das den Anspruch an Imputationsverfahren stellt, ebendiese bezüglich Verzerrungen und Vorhersagegenauigkeit evaluieren zu können. Des Weiteren hängt die Auswahl eines geeigneten Verfahrens direkt von den verwendeten Auswertungsmetriken ab, da die Beurteilung von Imputationsverfahren maßgeblich von den gewählten Metriken abhängt. Daher werden die Auswertungsmetriken im Folgenden zunächst anhand der identifizierten Studien qualitativ herausgearbeitet. In diesem Zuge werden sinnvolle Kategorien zur Einteilung der Auswertungsmetriken vorgestellt, anhand derer die einzelnen Metriken unterteilt werden können. Eine Zusammenfassung der Metriken in Kategorien hat den Vorteil, dass dadurch alle Auswertungsmetriken erfasst werden können. Auch wenn einzelnen Studien keine genauen Berechnungsformeln oder Einheiten zur Messung der Leistungsfähigkeit der Verfahren aufführen, können diese trotzdem eingeteilt werden. Zudem sind die Auswertungsmetriken der identifizierten Studien ähnlich vielfältig wie auch schon die Studiendesigns, weswegen eine Zusammenfassung mehrerer Metriken der Übersichtlichkeit dient und eine genaue Vorstellung jeder einzelnen Metrik obsolet macht. Diese Zusammenfassung dient darüber hinaus als Vorbereitung, um die Auswertungsmetriken mit den vorherigen Erkenntnissen der systematischen Literaturrecherche zu den Studiendesigns miteinander zu verknüpfen. Daher ist die in Abbildung 3-4 vorgeschlagene Einteilung grundsätzlich eng mit den verschiedenen Studiendesigns zum Vergleich der Verfahren verbunden, da je nach Ablauf und Rahmenbedingungen der Studie verschiedene Bewertungskriterien bzw. Auswertungsmetriken zur Verfügung stehen.

Grundsätzlich lassen sich der Ausgangsdatensatz und der vervollständigte Datensatz bei jeder Studie in irgendeiner Art und Weise miteinander vergleichen. Unabhängig davon, ob der Ausgangsdatensatz vollständig ist oder nicht, kann das statistische Verhalten der Datensätze vor und nach der Imputation miteinander verglichen werden. Der Vergleich des statistischen Verhaltens beschreibt den paarweisen Vergleich aggregierter Werte der Datensätze. In den beleuchteten Studien verwendete Parameter sind der Mittelwert, die Standardabweichung und die Varianz der Merkmale. Des Weiteren lässt sich die Verteilung der Ausprägungen eines Merkmals oder die Korrelation zwischen verschiedenen Merkmalen zur Bewertung heranziehen. Diese Werte lassen sich weiterhin in Lage- und Streuungsparameter unterteilen.

Basieren Studien auf vollständigen Datensätzen, aus denen zunächst Werte entfernt und anschließend wieder imputiert werden, so ergibt sich die Möglichkeit zum direkten Vergleich beider vollständiger Datensätze. Es können also alle einzelnen Werte beider Datensätze paarweise miteinander verglichen werden. So lässt sich die Vorhersagegenauigkeit jedes einzelnen Wertes, jeder Variable oder des ganzen Datensatzes bewerten. Dazu stehen verschiedene Auswertungsmetriken zur Verfügung, die die Vorhersagegenauigkeit bzw. die Abweichung der Werte messbar und somit beurteilbar machen. Grundsätzlich lassen sich diese Messwerte in relative und absolute Maßzahlen unterscheiden.

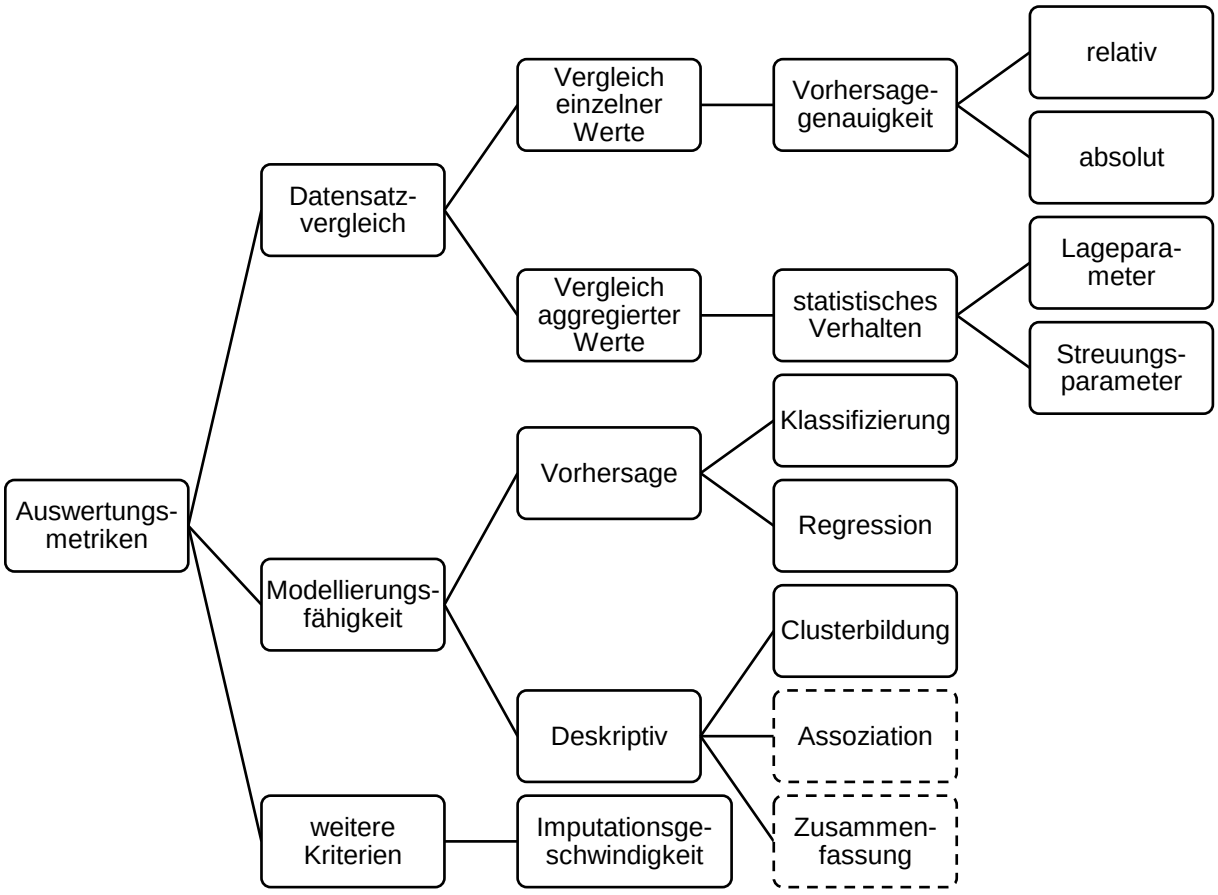

Abbildung 3-4: Auswertungsmetriken zum Vergleich von Imputationsverfahren

Sofern in einer Studie zusätzlich zur Imputation außerdem der weitergehende Schritt der Datenanalyse bzw. des Data Mining durchgeführt wird, so lässt sich neben einem Datensatzvergleich zusätzlich die Modellierungsfähigkeit des vervollständigten Datensatzes beurteilen. Die Modellierungsfähigkeit steht in dieser Arbeit für die Leistungsfähigkeit anschließender Data-Mining-Anwendungen. Wie im Grundlagenteil ausgeführt, existieren verschiedene Data-Mining-Aufgaben zu verschiedenen Analysezielen. Die Leistungsfähigkeit der entsprechenden Verfahren lässt sich mithilfe verschiedener Validierungstechniken beurteilen, die indirekt Aufschluss über die Datenvorverarbeitung und somit über die Qualität von Imputationsverfahren

bieten können. Wird beispielsweise ein Vorhersagemodell anhand der vervollständigten Daten entwickelt, so kann die Vorhersagegenauigkeit des Modells mithilfe einer Kreuz- oder Split-Validierung bewertet werden. Konkret bedeutet dies am Beispiel der Klassifizierung, dass anhand der vervollständigten Daten ein Modell entwickelt wird, anhand dessen ein Klassenattribut des Datensatzes reproduziert werden soll. Eine hohe Anzahl korrekt klassifizierter bzw. ein geringer Anteil fehlklassifizierter Objekte lässt im Umkehrschluss auf eine gute Eignung des verwendeten Imputationsverfahrens schließen. Dieses Vorgehen lässt sich analog dazu auch auf die anderen Data-Mining-Aufgaben anwenden, weswegen die Modellierungsfähigkeit in Abbildung 3-4 gemäß der in dieser Arbeit verwendeten Aufteilung der Data-Mining-Aufgaben unterteilt ist. Die Modellierungsfähigkeit lässt sich weitergehend ähnlich wie der Datensatzvergleich mithilfe von statistischen Kennzahlen beurteilen. Während sich bei der Clusteranalyse und bei der Klassifizierung die absolute oder relative Messung der korrekt eingeteilten Objekte anbietet, so lassen sich bei der Regressionsanalyse die relative und absolute Abweichung der vorhergesagten Werte messen. Die Data-Mining-Aufgaben der Zusammenfassung und der Assoziationsanalyse sind in Abbildung 3-4 in gestrichelten Konturen abgebildet, da sie in keiner der betrachten Studien herangezogen werden, sich aber theoretisch auch zur Auswertung verwenden ließen.

Darüber hinaus wird in einigen wenigen Studien die Imputationsgeschwindigkeit als Vergleichskriterium herangezogen. Die Imputationsgeschwindigkeit lässt sich durch die Dauer des jeweiligen Imputationsvorgangs charakterisieren und kann dadurch einen Anhaltspunkt zum Aufwand und zur Praktikabilität des jeweiligen Verfahrens darstellen. Besonders für spezielle Anwendungsfälle, die Echtzeit-Data-Mining erfordern, kann die Imputationsgeschwindigkeit ein entscheidendes Vergleichskriterium darstellen. Da darüber hinaus noch weitere Kriterien zur Praktikabilität denkbar wären, die jedoch nicht anhand der Studien identifiziert werden konnten, ist in Abbildung 3-4 ein Pfad für weitere Kriterien vorgesehen, dem auch die Imputationsgeschwindigkeit zugeordnet wird.

Rahmenkonzept zum Vergleich von Imputationsverfahren

In diesem Abschnitt werden die Ergebnisse der vorherigen Abschnitte übersichtlich dargestellt, indem ein Rahmenkonzept zum Vergleich von Imputationsverfahren vorgeschlagen wird. Die Idee dieses Rahmenkonzeptes ist es, alle Ergebnisse zum Vorgehen zum Vergleich von Imputationsverfahren übersichtlich zusammenzufassen. Das in Abbildung 3-5 dargestellte Rahmenkonzept aggregiert dazu die zuvor dargelegten Erkenntnisse zu den Studiendesigns, Rahmenbedingungen und Auswertungsmetriken.

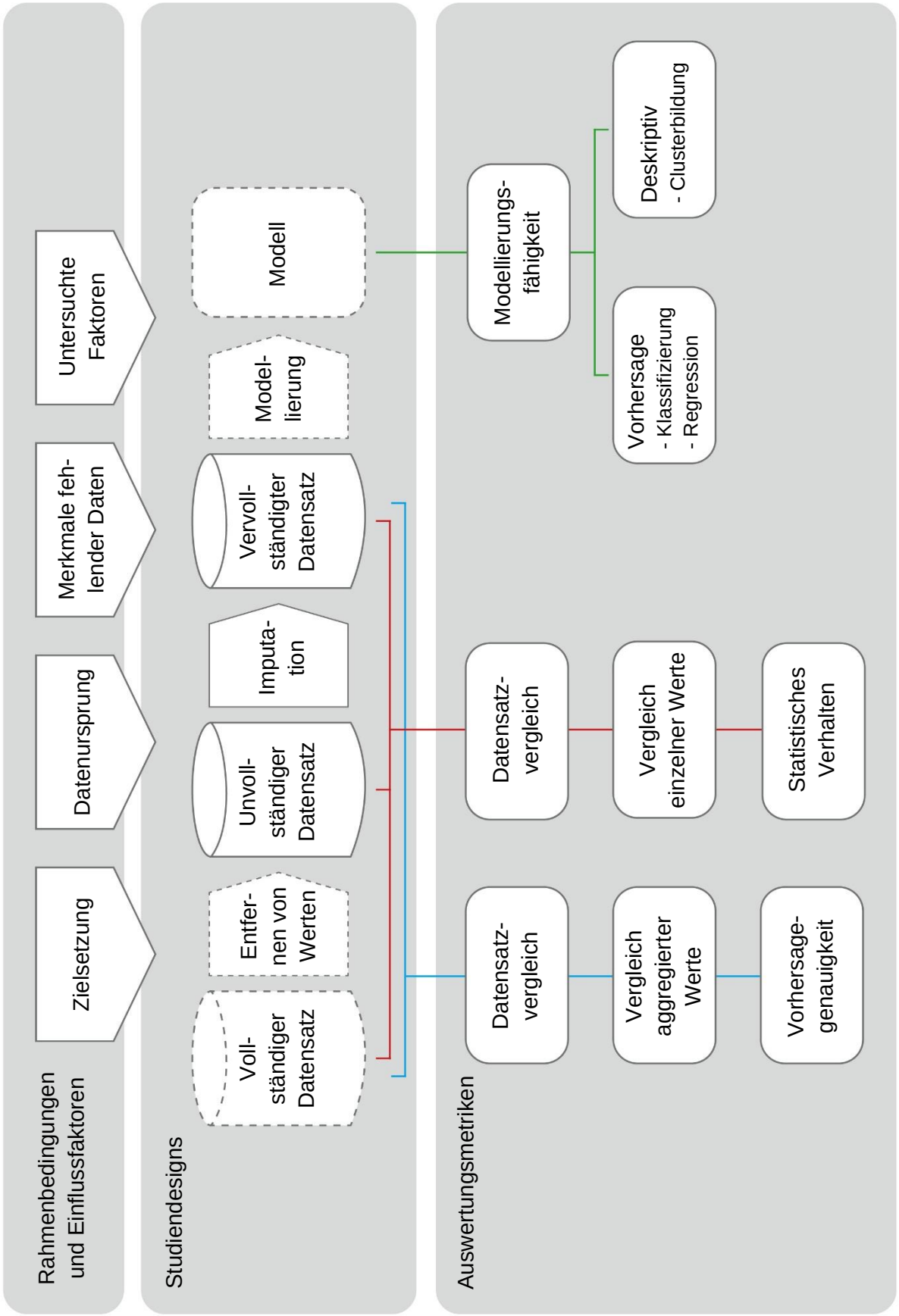

Abbildung 3-5: Rahmenkonzept zum Vergleich von Imputationsverfahren

Die Einflussfaktoren und Rahmenbedingungen, die sowohl das Studiendesign als auch die gewählten Auswertungsmetriken beeinflussen, sind dabei über den Ablaufelementen der Studien angeordnet. Die zuvor präsentierten Auswertungsmethoden und -metriken sind weitergehend direkt unter dem Studiendesign angeordnet, da sie in direkter Abhängigkeit zum jeweiligen Studiendesign stehen. Die blauen, grünen und roten Pfade verdeutlichen dabei, welche Auswertungsmetriken beim jeweiligen Studiendesign angewendet werden können. Ein Datensatzvergleich anhand des statistischen Verhaltens (roter Pfad in der Abbildung) kann sowohl bei unvollständigen als auch bei vollständigen Datensätzen als Bewertungskriterium zu Rate gezogen werden. Weitergehend kann auch die Modellierungsfähigkeit (grüner Pfad in der Abbildung) unabhängig von der Vollständigkeit der Daten als Bewertungskriterium verwendet werden. Ein Datensatzvergleich unter Bewertung der Vorhersagegenauigkeit (blauer Pfad in der Abbildung) ist logischerweise nur bei vollständigen Datensätzen als Grundlage möglich, da sich nur in diesem Fall die ursprünglichen Werte mit den anschließend entfernten und wiederum imputierten Werten vergleichen lassen.

3.2.2 Bewertung von Imputationsverfahren

Nachdem in Abschnitt 3.2.1 ausführlich erörtert wurde, wie sich Imputationsverfahren bewerten lassen, fokussiert dieser Abschnitt nun die Leistungsfähigkeit der Imputationsverfahren. Da das Ergebnis der Studien in der Regel durch die Leistungsfähigkeit bzw. eine Rangfolge der Imputationsverfahren in Abhängigkeit der Studiencharakteristika repräsentiert wird, werden dazu die Resultate der 78 identifizierten Studien qualitativ aggregiert. Konkret werden die Kernaussagen der Studien herausgearbeitet, miteinander verglichen und sofern möglich zusammengefasst. Anhand dieser Zusammenfassung werden anschließend möglichst allgemeingültige Handlungsempfehlungen zur Auswahl von Imputationsverfahren abgeleitet. Die vorangegangene Beschreibung der Studien anhand der aufgeführten Charakteristika kann dabei helfen, differenzierte Ergebnisse je Anwendungsfall herauszuarbeiten. Dabei kann die Zusammenfassung der Studienresultate auch Aufschluss darüber geben, inwiefern ebenjene Charakteristika die Ergebnisse und damit die Auswahl eines geeigneten Imputationsverfahrens beeinflussen können.

Bei der Betrachtung der Kernaussagen zur Beurteilung der verschiedenen Verfahren lassen sich auf den ersten Blick signifikante Unterschiede in der Bewertung zwischen Studien mit longitudinalen und nicht-longitudinalen Datensätzen ausmachen. Denn wie im zweiten Kapitel dargelegt, existieren spezielle Verfahren, die sich nur auf longitudinalen Datensätzen anwenden lassen. Daher wird in dieser Arbeit eine differenzierte Betrachtung der Studienergebnisse nach diesen beiden Datensatztypen angestellt.

Bei Studien mit nicht-longitudinalen Datensätzen lassen sich offensichtliche Unterschiede in der Leistungsfähigkeit bei traditionellen und modernen Methoden feststellen. In einem Großteil der Studien mit nicht-longitudinalen Datensätzen werden stellvertretende Techniken beider Gruppen getestet und gegenübergestellt. Erwartungsgemäß übertreffen die Ergebnisse der modernen Methoden in jeder dieser Studien die Resultate der traditionellen Verfahren. Ein Großteil dieser Studien, in denen weiterhin verschiedene Ausfallraten getestet werden, stellt zudem fest, dass die Unterschiede in der Imputationsqualität gemessen an den jeweils verwendeten Auswertungsmetriken mit steigender Ausfallrate zunehmen. Im Umkehrschluss bedeutet dies, dass sich die Ergebnisse bzw. die Leistungsfähigkeiten der verschiedenen Verfahren bei niedrigen Ausfallraten weniger stark voneinander unterscheiden. Verschiedene Studien geben diesbezüglich an, dass einfache respektive traditionelle Verfahren bei Ausfallraten bis zu 10% bzw. 20% unabhängig von anderen Einflussfaktoren nahezu gleichwertige Ergebnisse wie fortgeschrittene Verfahren liefern können. Einzelne Studien zeigen weitergehend auf, dass sich die Rangfolge innerhalb dieser Spanne bei Veränderung der Ausfallrate häufig verändert, wohingegen sich mit steigender Ausfallrate in der Regel eine feste Rangfolge der Verfahren abzeichnet. Dies wird auch dadurch verdeutlicht, dass die qualitativen Unterschiede zwischen den Verfahren bei steigender Ausfallrate ebenfalls wachsen. Nachdem bereits dargelegt wurde, dass fortgeschrittene Methoden in der Regel die Ergebnisse von einfachen Methoden übertreffen, werden im Folgenden diejenigen Studien betrachtet, die mindestens zwei moderne Verfahren gegenüberstellen. Hier ist direkt ersichtlich, dass die Ergebnisse der Studien ähnlich vielfältig wie die die Studiendesigns sind. Nahezu jedes Verfahren bzw. jede Verfahrensgruppe konnte in einzelnen Anwendungsfällen am meisten überzeugen und die Ergebnisse anderer Verfahren übertreffen. Multiple Imputationsverfahren, Imputationsverfahren auf Grundlage von angepassten Regressionsmodellen, die ML-Estimation und verschiedene Imputationsverfahren auf Basis von Entscheidungsbaummodellen können demnach für verschiedene Anwendungsfälle die beste Option darstellen. Außerdem werden in diversen domänenspezifischen Veröffentlichungen domänenspezifische Imputationsverfahren oder Modelle zur Imputation verwendet, die an dieser Stelle nicht in den Vergleich mit eingehen, da sie keinen Mehrwert für die Datenwissenschaft darbieten.

Innerhalb der Multiplen Imputationsverfahren geht ein Großteil der Veröffentlichungen in der Aussage überein, dass der JM-Ansatz respektive die MVNI grundsätzlich mindestens gleichwertige Ergebnisse wie der FCS-Ansatz liefert. Allerdings lässt sich die letztgenannte Methode demnach einfacher implementieren und stellt im Aufwand-Nutzen-Vergleich die bessere Alternative dar. Außerdem sind sich alle Veröffentlichungen, die multiple Imputationsverfahren miteinander vergleichen, einig, dass der volle Mehrwert dieser Imputationsverfahren nur bei anwendungsfallspezifischer Anpassung ausgeschöpft werden kann. Werden standardmäßige

Softwareimplementierungen der multiplen Imputation anderen modellbasierten Ansätzen gegenübergestellt, so können diese selten einen oberen Platz in der Rangfolge der Imputationsverfahren belegen. Multiple Imputationsverfahren sind also gerade bei genauer Auseinandersetzung mit dem Anwendungsfall und bei entsprechender Parametrisierung durchaus zu empfehlen. Außerdem können Multiple Imputationsverfahren durch die Möglichkeit der Kombination mit anderen Modellen für jeden Anwendungsfall sehr gute oder sogar die besten Ergebnisse liefern. Weitergehend bleibt zu beachten, dass auch die standardmäßigen, unangepassten Implementationen der multiplen Imputationsverfahren bei nicht-longitudinalen Daten ab einer gewissen Ausfallrate in der Regel allen traditionellen Methoden überlegen sind. Ist der Datensatz bzw. der spezifische Anwendungsfall allerdings unbekannt oder wird im Vorfeld nicht genauer beleuchtet, wodurch folglich keine Anpassung der multiplen Imputation möglich ist, sollte gegebenenfalls die Verwendung anderer fortgeschrittener Methoden erwägt werden.

In den Studien verglichene Implementationen auf Grundlage von Entscheidungsbaummodellen (Classification and Regression Trees (CART), Random Forest (RF), missForest, XGBoost) stellen bei der Gegenüberstellung mit anderen fortgeschrittenen Methoden nur selten die beste Option dar. Allerdings gehen die diesbezüglichen Studienresultate in der These einher, dass Entscheidungsbaumverfahren unabhängig vom Anwendungsfall und ohne großen Aufwand zumeist sehr gute Ergebnisse liefern, ohne dabei alle anderen Verfahren zu übertreffen. Implementationen auf Basis von Entscheidungsbäumen können dabei nach einigen Veröffentlichungen insbesondere bei Datensätzen mit kategorialen bzw. nicht-kontinuierlichen Merkmalen sehr gute oder gar die besten Resultate liefern. Andere Veröffentlichungen legen nahe, dass die einfache Implementierung und anwendungsfallunabhängig gute Ergebnisse klare Argumente für die Auswahl von Entscheidungsbaumverfahren darstellen. Entscheidungsbaumverfahren bzw. -modelle zur Imputation sind demnach besonders dann zu empfehlen, wenn der Imputation innerhalb eines Wissensentdeckungsprozesses kein zu großer Aufwand zuteilwerden soll, aber gleichzeitig gute Ergebnisse erforderlich sind. Außerdem kann ein hoher Anteil kategorialer Merkmale innerhalb eines Datensatzes für die Verwendung von Entscheidungsbaummodellen zur Imputation sprechen.

Die Ergebnisse zu regressionsbasierten Methoden und zur ML-Estimation lassen sich noch schwieriger verallgemeinern. Regressionsmodelle können, wie im Grundlagenteil ausgeführt, auf vielfältige Weise angepasst und angewendet werden, weswegen sich die Ergebnisse und die Platzierungen in der Rangfolge der Imputationsverfahren zwischen den Studien stark unterscheiden. Zudem werden Regressionsmodelle häufig innerhalb der multiplen Imputation verwendet, was eine genaue Abgrenzung dieser Methoden erschwert. Ähnlich wie bei den multiplen Imputationsverfahren kann zur Regressionsimputation geschlussfolgert werden,

dass die Verwendung von angepassten Regressionsmodellen in der Regel sehr gute Ergebnisse liefert und die Auseinandersetzung mit dem Anwendungsfall bzw. ein spezifischer Vergleich von Verfahren daher essenziell ist. Die ML-Estimation wird, obwohl es sich in der Theorie um ein state-of-the-art-Verfahren handelt, eher selten in Vergleiche miteinbezogen. In diesen wenigen Vergleichen kann die ML-Estimation in kaum einer Studie einen der obersten Plätze in der Rangfolge der Imputationsverfahren belegen, ist jedoch auf der anderen Seite einfach zu implementieren und setzt in der Regel unabhängig vom Anwendungsfall akzeptable bzw. plausible Werte ein. Insbesondere die Verwendung der ML-Estimation als Modell für die Multiple Imputation kann sehr gute Ergebnisse erzielen und liefert in einer der Studien mit anschließendem Data Mining im Durchschnitt die besten Ergebnisse über alle Data-Mining-Verfahren hinweg.

Die Studien mit longitudinalen Datensätzen können die zuvor aufgeführten Beobachtungen nicht uneingeschränkt bestätigen. Zwar lassen sich die Erkenntnisse bezüglich des Einflusses der Ausfallrate gemäß den meisten Studien auch auf longitudinale Daten übertragen, allerdings gibt es Unterschiede in der Bewertung der Imputationsverfahren. Einige Studien zeigen auf, dass einzelne, naive Verfahren, wie z. B. das Einsetzen des vorangegangenen Wertes oder die Verwendung des Mittelwertes der vorherigen Werte, in bestimmten Anwendungsfällen mit fortgeschrittenen, modellbasierten Verfahren konkurrieren können. Außerdem zeigen nahezu alle Studien mit longitudinalen Daten, dass speziell angepasste Verfahren für diesen Datensatztypen die Ergebnisse der allgemein anwendbaren Verfahren übertreffen. In den Veröffentlichungen wird begründet, dass spezielle Verfahren für longitudinale Datensätze die zeitliche Abhängigkeit der Merkmalsausprägungen mit einbeziehen und demzufolge bessere Vorhersagen für diesen Datensatztypen hervorbringen können.

Besonders interessant für den Kontext dieser Arbeit ist die Kombination aus verschiedenen Imputations- und Data-Mining-Verfahren. Deshalb werden die 19 Veröffentlichungen, die den zusätzlichen Schritt der Modellierung beinhalten, an dieser Stelle noch einmal gesondert beleuchtet. Zwar können anhand dieser 19 Studien keine allgemeingültigen Aussagen über die Rangfolge einzelner Verfahren als Vorbereitung für Data Mining abgeleitet werden, allerdings können diese Studien beispielsweise Aufschluss darüber geben, inwiefern die Imputationsqualität mit der anschließenden Modellierungsfähigkeit zusammenhängt. Diejenigen Veröffentlichungen, die verschiedene Imputationsverfahren mit einem einzigen Data-Mining-Verfahren kombinieren, legen diesbezüglich nahe, dass die Imputationsqualität gemessen an der Vorhersagegenauigkeit mit der Modellierungsfähigkeit einhergeht. Einige dieser Veröffentlichungen gehen sogar weiter, indem sie anhand der Studienergebnisse schlussfolgern, dass Unterschiede in der Vorhersagegenauigkeit in der Modellierungsfähigkeit verstärkt abgebildet werden. Aus dieser Tatsache kann abgeleitet werden, dass auch marginale Unterschiede in

der Imputationsqualität, die gemäß den vorherigen Ausführungen bei vergleichsweise niedrigen Ausfallraten vorkommen, zu deutlichen Unterschieden in den Ergebnissen nach dem Data Mining führen. Auf der anderen Seite zeigen diejenigen Studien, die Kombinationen aus verschiedenen Imputationsmethoden und verschiedenen Data-Mining-Verfahren gegenüberstellen, teilweise auch, dass je nach Verfahrenskombination verschieden gute Ergebnisse erzielt werden können. Diesen Studien zufolge können Imputationsverfahren mit einer geringeren Vorhersagegenauigkeit unter Verwendung eines geeigneten Data-Mining-Verfahrens durchaus gute Ergebnisse für Wissensentdeckungsprozesse in Datenbanken liefern.

3.3 Diskussion der systematischen Literaturrecherche

Die betrachteten Studien und deren Aggregation können zweifelsohne gute Anhaltspunkte zum Vergleich bzw. zur Bewertung von Imputationsverfahren bieten. Besonders anhand der Vorgehensweisen und der damit verbundenen Auswertungsmetriken können Handlungsempfehlungen für zukünftige Vergleiche von Imputationsverfahren abgeleitet werden.

Allerdings deckt diese Arbeit, die Imputationsverfahren insbesondere im Kontext der Wissensentdeckung und in Bezug auf Data Mining betrachtet, gerade an dieser Stelle gewisse Forschungslücken auf. Nur etwa ein Viertel der Vergleichsstudien stellt überhaupt einen Bezug zu Datenanalyse- oder Data-Mining-verfahren her. Noch weniger Studien beleuchten Imputationsverfahren explizit im Kontext eines Wissensentdeckungsprozesses in Datenbanken. Weitergehend bedeutet das, dass eine Bewertung, inwiefern sich bestimmte Kombinationen aus Imputationsverfahren und Data-Mining-Aufgabe besser oder schlechter für Wissensentdeckungsprozesse in Datenbanken eignen, anhand der identifizierten Veröffentlichungen nur schwierig zu verallgemeinern ist.

Da insbesondere die Vorhersagegenauigkeit Aufschluss über die Datenqualität geben kann, ist die Vorhersagegenauigkeit das wohl wichtigste Bewertungskriterium zum Vergleich von Imputationsverfahren. Daher kann auch angenommen werden, dass die Vorhersagegenauigkeit indirekt auch Modellierungsfähigkeit anschließender Data-Mining-Prozesse einhergeht.

Zwar erscheint die Bewertung des statistischen Verhaltens bei der Verwendung vollständiger Datensätze im Vergleich zur Bewertung der Vorhersagegenauigkeit als weniger aussagekräftig und hinfällig, allerdings kann das statistische Verhalten als zusätzliches Kriterium herangezogen werden. Besonders bei Studien, denen ein unvollständiger Datensatz zugrunde liegt, aus dem aber nur die vollständigen Objekte für die eigentliche Studie verwendet werden, könnte das statistische Verhalten aufschlussreiche Informationen liefern. Die ausschließliche

Verwendung des statistischen Verhaltens, obwohl ein ursprünglich vollständiger und ein vervollständigter Datensatz vorliegen, ist keinesfalls zu empfehlen, da objektiv bessere Bewertungskriterien zum Vergleich der Imputationsqualität zur Verfügung stehen.

Da nur in einzelnen Fällen unvollständige Realdaten verwendet werden, bleibt die Frage bestehen, wie sich die Eignung von Imputationsverfahren bewerten lässt, sofern diese auf Realdaten mit fehlenden Werten angewendet werden. Dabei kann die Modellierungsfähigkeit als Bewertungskriterium durchaus Abhilfe schaffen. Die Modellierungsfähigkeit eignet sich insbesondere dann als Beurteilungskriterium, wenn die Zielvariable der Datenanalyse, z. B. in Form eines Klassenattributs, in realen, unvollständigen Datensätzen vollständig beobachtet ist. Sofern die Modellierungsfähigkeit bzw. die Leistungsfähigkeit des Data Mining im Nachgang von Imputationsverfahren verbessert werden kann, so kann darüber auf die Qualität der Imputationsverfahren und -ergebnisse geschlossen werden.

Die Aggregation der Leistungsfähigkeit einzelner Verfahren auf Grundlage vorhandener Studien stellt eine weitaus größere Herausforderung als die Synthese der Studiendesigns dar. Gerade wegen der ausführlich dargelegten Vielfältigkeit der Studiendesigns ist eine allgemein gültige und objektive Bewertung auf Grundlage der betrachteten Veröffentlichungen nahezu unmöglich. Vielmehr erscheint es für die Ableitung von Handlungsempfehlungen und für die Auswahl eines geeigneten Verfahrens sinnvoll, diejenigen Studien heranzuziehen, die dem jeweiligen Anwendungsfall am ähnlichsten sind. Genau an dieser stellt die hier angestellte systematische Literaturrecherche einen Mehrwert dar. Durch die vorherige Charakterisierung der einzelnen Studien anhand der Rahmenbedingungen, Einfluss- und Untersuchungsfaktoren lassen sich nämlich für nahezu jeden möglichen Anwendungsfall vergleichbare Studien mit dazugehörigen Resultaten identifizieren.

Eine allgemeine Bewertung der Verfahren mithilfe der systematischen Literaturrecherche ist daher nur eingeschränkt möglich. Vielmehr ist die Auswahl eines geeigneten Imputationsverfahrens weiterhin eine vom Anwendungsfall abhängige Entscheidung, bei der diese systematische Literaturrecherche allerdings als entscheidende Hilfestellung dienen kann. Dabei kann die Auswertung der Vorgehensweisen in Kombination mit der Bewertung der Verfahren Aufschluss über den Einfluss der verschiedenen Rahmenbedingungen geben. Nach Betrachtung der Ergebnisse ist eine differenzierte Betrachtung der Studienergebnisse nach longitudinalen und nicht-longitudinalen Datensätzen in jedem Fall sinnvoll, da die Leistungsfähigkeit der Verfahren sich je nach Datensatztyp teilweise stark unterscheidet. Eine differenzierte Betrachtung nach den anderen Einflussfaktoren und Rahmenbedingungen mit Ausnahme der Ausfallrate erscheint nach Betrachtung der Studienresultate weniger bedeutend. Viele Studien zeigen,

dass sich die Bewertung bzw. die Rangfolge der Verfahren nicht maßgeblich durch die Veränderung der anderen aufgeführten Faktoren verändert. Zudem zeigen nahezu alle Studien, die mehrere Auswertungsmetriken zur Bewertung der Imputationsqualität heranziehen, dass die Bewertung anhand verschiedener Auswertungsmetriken zumeist dieselbe oder eine zumindest ähnliche Rangfolge der Verfahren ergibt. In keiner der Studien konnte ein Verfahren anhand eines Kriteriums also gänzlich überzeugen und gleichzeitig anhand eines anderen Kriteriums signifikant schlechtere Ergebnisse liefern.

Obwohl viele Studien keine Veränderung der Rangfolge der Imputationsverfahren in Abhängigkeit von bestimmten Rahmenbedingungen feststellen, sprechen die Ergebnisse der sonst teilweise widersprüchlichen Studienergebnisse dafür, dass die Auswahl eines geeigneten Verfahrens immer dem konkreten Anwendungsfall unterliegt. Daher ist für optimale Ergebnisse eine sorgfältige Auswahl und gegebenenfalls ein vom Anwendungsfall abhängiger Vergleich der Verfahren empfehlenswert. Besonders bei Ausfallraten ab etwa 20% ist die Auswahl eines Imputationsverfahrens definitiv entscheidend und wirkt sich auf die Imputationsqualität und damit in vielen Fällen auch auf die Modellierungsfähigkeit aus. Ist die Ausfallrate niedriger, der Anwendungsfall unbekannt oder der Imputation wird innerhalb eines Wissensentdeckungsprozesses kein zu großer Anteil zugerechnet, so empfiehlt es sich, den Nutzen in Relation zum Aufwand zu betrachten.

Da die Resultate der Studien mit nicht-longitudinalen Datensätzen nahelegen, dass auch viele moderne und modellbasierte Methoden einfach zu implementieren sind, ist es nach der systematischen Literaturrecherche ratsam, zumindest die standardmäßigen Ausführungen ebendieser Methoden zu verwenden. Einfache Methoden wie die Mittelwertsimputation, die Zufallszahlimputation, Deck-Methoden und auch Imputationsmethoden auf Basis von Distanzeigenschaften können nur in einzelnen, speziellen Anwendungsfällen oder nach Anpassung der jeweiligen Methode überzeugen. Daher sind die anderen modellbasierten und modernen Verfahren bei nicht-longitudinalen Datensätzen grundsätzlich zu bevorzugen. Bei ausführlicher Auseinandersetzung mit dem konkreten Anwendungsfall ist die Verwendung von speziell angepassten bzw. parametrisierten Verfahren empfehlenswert. Hier eignen sich insbesondere multiple Imputationsverfahren unter Verwendung bestimmter Modelle und angepasste Regressionsmodelle bzw. die Kombination aus diesen Verfahren zur Imputation. Die in den Veröffentlichungen getesteten Entscheidungsbaummodelle zur Imputation liefern unabhängig vom Anwendungsfall konstant gute, aber nicht unbedingt die besten Ergebnisse. Ähnlich verhält es sich beim Maximum-Likelihood-Ansatz, dessen voller Mehrwert aber wohl erst im Rahmen der multiplen Imputation ausgeschöpft werden kann. Eine allgemeine Bewertung weiterer

modellbasierter Verfahren kann aufgrund der zu kleinen Stichproben nicht durchgeführt werden, wobei anzumerken ist, dass modellbasierte Verfahren in nahezu allen Studien den zuvor aufgeführten, einfachen Methoden überlegen sind.

Bei longitudinalen Daten können spezifische Methoden für ebenjenen Datensatztypen in der Regel bessere Ergebnisse hervorbringen als allgemein anwendbare Methoden. Sogar naive Methoden wie das Einsetzen des vorangegangenen Wertes oder das Einsetzen des Mittelwertes der vorangegangenen Werte können häufig bessere Ergebnisse als fortgeschrittene, modellbasierte Methoden liefern.

Eine objektive Bewertung speziell im Kontext von Wissensentdeckungsprozessen bzw. eine Bewertung der Eignung als Vorbereitung für anschließende Data-Mining-Verfahren sind aufgrund des zu kleinen Stichprobenumfangs noch schwieriger anzustellen. Die in dieser Arbeit aufgeführten Ergebnisse basieren hier jeweils auf einzelnen Studien, deren Übertragbarkeit nicht gewährleistet ist. Außerdem widersprechen sich die Studien bzw. deren Ergebnisse insbesondere dann, wenn Schlussfolgerungen bezüglich Abhängigkeiten zwischen Imputationsqualität und Modellierungsfähigkeit gezogen werden.

Neben der Aggregation und Auswertung der Ergebnisse lässt die systematische Literaturrecherche außerdem eine Beurteilung der Qualität der Veröffentlichungen selbst zu. Das kann dazu beitragen, mögliche Forschungslücken oder interessante Fokuspunkte für die zukünftige Forschungen aufzudecken. Diesbezüglich ist beim Aufbau der Veröffentlichungen und beim Vorgehen der jeweiligen Studien auffällig, dass jede Studie insofern einzigartig ist, dass sie einem eigenen, an das Anwendungsbeispiel angepasste Vorgehen folgt. Dadurch findet in den Studien ein Vergleich auf unterschiedlichen Ebenen statt. In einigen Studien werden übergeordnete Verfahrensgruppen miteinander verglichen, indem stellvertretende Techniken der jeweiligen Verfahrensgruppe gegenübergestellt werden. Andere Studien vergleichen direkt konkrete Verfahren oder gar Algorithmen und Software-Implementationen miteinander, wobei teilweise auch diverse Techniken derselben Verfahrensgruppe untereinander konkurrieren. Die Ergebnisse der verschiedenen Studien unterscheiden sich dementsprechend darin, dass die erste Art von Studien eine Bewertung der übergeordneten Verfahrensgruppen und die zweite Art von Studien eine detailliertere Bewertung konkreter Techniken zulässt. Daraus folgen Herausforderungen bei der Synthese der Ergebnisse und bei der Ableitung und Übertragung der Forschungsergebnisse.

Darüber hinaus ist auffällig, dass kaum eine Veröffentlichung über die Beschreibung der allgemeinen Methodik der Studie hinausgeht, indem die eigentliche Durchführung der Studie und insbesondere der Imputationsverfahren im Detail beschrieben wird. Die verwendete Software

oder Programmiersprache wird zwar in einigen Fällen erwähnt, allerdings in keiner der identifizierten Studien näher ausgeführt, indem das Vorgehen innerhalb der Software oder der verwendete Programmcode in der Veröffentlichung präsentiert werden. Diese Tatsache führt dazu, dass die Studien zwar Aufschluss darüber geben können, welche Verfahren sich besonders gut oder schlecht für bestimmte Anwendungen eignen, allerdings nicht darüber, wie diese sich implementieren lassen.

Zudem erscheint die Auswahl der getesteten Verfahren und Techniken in einem Großteil der Studien willkürlich, da die Auswahl nicht explizit begründet wird. Diejenigen Studien, in denen die Verfahrensauswahl thematisiert wird, verfolgen zumeist das Ziel, einem herkömmlichen Verfahren ein verbessertes Verfahren gegenüberzustellen, was die Auswahl in diesem Fall implizit begründet. Aus dieser Tatsache folgt, dass aus den betrachteten Studien keine Methodik oder ähnliches zur Auswahl eines geeigneten Verfahrens abgeleitet werden kann. Das vorgeschlagene Rahmenkonzept zeigt zwar auf, wie verschiedene Verfahren miteinander verglichen werden können, allerdings unterliegt die Auswahl der jeweils verwendeten Elemente noch immer dem Anwender.

Die gerade aufgeführten Probleme der betrachteten Veröffentlichungen haben zur Folge, dass eine objektive Bewertung von Imputationsverfahren auf Grundlage von Literaturergebnissen äußerst schwierig zu realisieren ist. Daher gehen auch eine Vielzahl der Veröffentlichungen in der Charakteristik einher, dass sie zögerlich in der Ableitung von Handlungsempfehlungen sind. Die meisten Veröffentlichungen implizieren, dass die Ergebnisse der jeweiligen Studie nicht oder nur eingeschränkt übertragbar sind.

Abschließend lässt sich nach der systematischen Literaturrecherche das Fazit ziehen, dass Imputationsverfahren zwar bereits auf vielfältige Art und Weise erforscht und verglichen wurden, dass sich trotzdem oder gerade deswegen nur wenig allgemein gültige Schlussfolgerungen ziehen lassen. Insbesondere die Untersuchungen von Imputationsverfahren in Verbindung mit anschließendem Data Mining lassen noch einige Fragen offen. Daher sind gerade diesbezüglich weitere Studien erforderlich. Damit lässt sich aufgrund der systematischen Literaturrecherche das Fazit ziehen, dass die Auswahl eines geeigneten Imputationsverfahren eine einzelfallabhängige Entscheidung bleibt, bei der die vorherigen Ausführungen aber durchaus behilflich sein können.

4 Exemplarische Anwendung von Imputationsverfahren

Die folgenden Abschnitte beschäftigen sich schließlich mit der exemplarischen Anwendung verschiedener Imputationsverfahren als Vorbereitung für anschließendes Data Mining im Kontext eines Wissensentdeckungsprozesses in Datenbanken. Dieses Kapitel greift dazu einige Erkenntnisse des dritten Kapitels auf und versucht gleichzeitig, einige der in der Literaturrecherche aufgedeckten Forschungslücken zu schließen. Dazu werden verschiedene Imputationsverfahren als Vorbereitung für Data Mining anhand eines konkreten Fallbeispiels angewendet. Für diese exemplarische Anwendung werden die wesentlichen Ablaufphasen, Aufgaben und Aktivitäten des CRISP-DM-Vorgehensmodell durchgeführt und in den folgenden Abschnitten beschrieben. Dabei stehen der Umgang mit fehlenden Merkmalswerten und insbesondere die Durchführung der Imputationsverfahren im Mittelpunkt. Neben dem reinen Vergleich der Verfahren stehen insbesondere die genaue Beschreibung der Anwendung und die Kontextualisierung im Vordergrund.

4.1 Auswahl des Fallbeispiels und Vorgehen zur exemplarischen Anwendung

Da die gänzlich neue Entwicklung eines KDD-Prozesses samt Datensammlung und Formulierung des Analyseproblems den Umfang dieser Arbeit übersteigen würde, wird auf ein bereits erforschtes Anwendungsbeispiel aufgebaut. Die in diesem Kapitel vorgenommene exemplarische Anwendung basiert auf einer Ausschreibung des Unternehmens Scania im Rahmen der *Industrial Challenge 2016 at the 15th International Symposium on Intelligent Data Analysis (IDA 2016).* Bei der Suche nach geeigneten Datensätzen für die Anwendung in dieser Arbeit fiel die Auswahl auf dieses konkrete Fallbeispiel, da es für diese Arbeit gleich mehrere Vorteile bietet. Die Ausschreibung enthält mehrere Datensätze sowie eine dazugehörige Beschreibung mitsamt einer impliziten Aufgabenstellung. Es handelt sich bei den veröffentlichten Datensätzen um reale, in der Praxis aufgenommene Daten eines großen Unternehmens aus dem produzierenden Sektor. Wie die systematische Literaturrecherche gezeigt hat, werden nur äußerst selten reale Datensätze dieser Domäne zur Untersuchung verwendet, weswegen in dieser Domäne häufig auf Erkenntnisse anderer Anwendungsbereiche zurückgegriffen werden muss. Weitergehend sind die Daten selbst von vorneherein unvollständig, was zwangsläufig einen Umgang mit fehlenden Merkmalswerten erfordert. Zudem erlaubt die Veröffentlichung im Kontext eines Wettbewerbs bzw. einer Konferenz die freie Verwendung für wissenschaftliche Arbeiten.

Wie bereits in der Einleitung ausgeführt und zusätzlich im dritten Kapitel dieser Arbeit gezeigt, werden Imputationsverfahren häufig losgelöst von Wissensentdeckungsprozessen und anschließenden Data-Mining-Verfahren betrachtet und erforscht. Darüber hinaus werden häufig simulierte Daten zur theoretischen Untersuchung von Imputationsverfahren genutzt, da die Variation von Objektanzahl, Attributsanzahl und Korrelation zwischen den Merkmalen die Untersuchung einzelner Einflussfaktoren zulässt. Die hier ausgeführte exemplarische Anwendung verfolgt im Gegensatz zum Großteil der zuvor betrachteten Studien explizit das Ziel, Imputationsverfahren und deren Eignung im Kontext ganzer Wissensentdeckungsprozesse in Datenbanken zu beleuchten. Dazu wird die Leistungsfähigkeit des Wissensentdeckungsprozesses bei Verwendung verschiedener Imputationsverfahren untersucht. Konkret werden bestimmte Elemente des zuvor vorgestellten Rahmenkonzepts zur Bewertung von Imputationsverfahren in einen Wissensentdeckungsprozess nach CRISP-DM eingebunden. Dazu werden die Aufgaben und Aktivitäten der Datenvorverarbeitung sowie die der Modellierung und Auswertung bei sonst gleichbleibenden Bedingungen unter Veränderung der Datensätze, der Imputationsverfahren und Data-Mining-Verfahren wiederholt durchgeführt. Weitergehend werden beide Ansätze zur Untersuchung von Imputationsverfahren, basierend auf vollständigen und unvollständigen Datensätzen, angewendet. Die Auswertung findet dabei auf Grundlage der Modellierungsfähigkeit statt, da bei dieser Anwendung explizit die Leistungsfähigkeit des gesamten Wissensentdeckungsprozesses in Abhängigkeit von Imputationsverfahren und eben nicht die reine Imputationsqualität im Vordergrund steht. Bezogen auf das vorgestellte Rahmenkonzept werden also sowohl per se unvollständige Daten als auch durch das Entfernen von Werten unvollständige Datensätze mithilfe von Imputationsverfahren vervollständigt und anschließend zur Modellierung verwendet. Die Modellierungsfähigkeit stellt demzufolge das Bewertungskriterium dieser exemplarischen Anwendung zum Vergleich von Imputationsverfahren dar.

4.2 Durchführung anhand des Fallbeispiels

Die folgenden Abschnitte beschreiben schließlich den genauen Ablauf und die Durchführung anhand des ausgewählten Fallbeispiels. Damit präsentieren die folgenden Absätze einen Ablauf zum Vergleich von Imputationsverfahren im Kontext von Wissensentdeckungsprozessen, der sich auch auf andere Fallbeispiele übertragen lässt. Da die Durchführung von KDD-Prozessen sowohl die Schritte zur Anwendung als auch zur Auswertung enthalten, sind die Ausführungen gemäß den CRISP-DM-Phasen unterteilt. Während unterstützende Aufgaben des Wissensentdeckungsprozesses, wie die Aufgaben der Phasen Domänen- und Datenverständnis, in kurzer Berichtsform präsentiert werden, werden die Aufgaben, die direkt anhand der Daten geschehen, ausführlicher dargestellt. Die letztgenannten Aufgaben werden mithilfe der

Datenanalysesoftware RapidMiner Studio in der Version 10.1 (nachfolgen RapidMiner genannt) umgesetzt und auch anhand dessen erklärt. Einzelne Programmbausteine, die nicht standardmäßig in RapidMiner enthalten sind, werden weitergehend mithilfe von Programmcodes der Programmiersprache Python in der Version 3.9.13 implementiert.

Der Abschnitt 4.2.1 zum Domänen- und Datenverständnis fasst relevante Informationen zum Ursprung der Daten, zum Datensatz selbst und zum Ziel des KDD-Prozesses zusammen. Die Ausführungen zur Datenvorverarbeitung in Abschnitt 4.2.2 gehen kurz auf die zur Durchführung notwendiger Vorverarbeitungsschritte und ausführlich auf die Imputation fehlender Daten und deren Integration in den Gesamtablauf ein. Abschnitt 4.2.3 beschreibt die Schritte zur Modellierung bzw. zum Data Mining und zur Auswertung der Ergebnisse. Die Ergebnisse selbst werden in Abschnitt 4.2.4 dargestellt.

4.2.1 Domänenverständnis und Datenverständnis

Die ersten beiden Phasen des CRISP-DM-Vorgehensmodells bestehen aus unterstützenden bzw. vorbereitenden Aufgaben für den weiteren Ablauf. Dabei bleiben die verwendeten Daten selbst noch unberührt. Vielmehr geht es darum, ein Verständnis für das betrachtete Geschäftsfeld, die Daten und den Zweck des Prozesses zu erlangen. Wesentliche Aufgaben der Phase Business Understanding bestehen im Festhalten der geschäftlichen Ziele und der damit verbundenen Aufgabendefinition. Das Data Understanding umfasst gegebenenfalls die Sammlung und insbesondere die erste Sichtung der Daten. Da diese beiden Phasen nicht im Fokus dieser Arbeit stehen, aber dennoch wichtig für das Verständnis des Fallbeispiels sind, werden die wesentlichen Aspekte in diesem Abschnitt zusammengefasst, ohne dabei alle Aufgaben im Einzelnen abzuhandeln. Außerdem sind bereits viele Punkte dieser beiden Phasen durch die Ausschreibung des Anwendungsbeispiels abgedeckt und werden daher nicht gänzlich neu erarbeitet, sondern lediglich strukturiert ausformuliert.

Die für diese Studie verwendeten Daten und deren Beschreibung stammen von dem schwedischen Lastkraftwagenhersteller Scania. Konkret handelt es sich um Massendaten, die im alltäglichen Gebrauch von Scania-Fahrzeugen gesammelt wurden. Bei der Datenanalyse steht das sogenannte *Air Pressure System (APS)* im Fokus, das wichtige Fahrzeugkomponenten wie die Gangschaltung oder die Bremsen mit Druckluft versorgt.

Die aufgenommenen Daten zu den jeweiligen Fahrzeugen sind dabei in einen Trainingsdatensatz mit insgesamt 60.000 Objekten und einen Testdatensatz mit insgesamt 16.000 Objekten unterteilt, wobei jedes Objekt ein Fahrzeug repräsentiert. Weitergehend wird jedes Objekt durch 171 Attribute beschrieben. Ein klassifizierendes Attribut gibt an, ob das APS tatsächlich

fehlerbehaftet ist, während alle anderen Merkmale durch numerische Werte beschrieben werden, die dem alltäglichen Gebrauch der Fahrzeuge entstammen. Dabei sind die ursprünglichen Attributsbezeichnungen und damit auch die Einheiten aller Werte aus privaten Gründen des Herausgebers anonymisiert. Eine Anonymisierung hat für die Verwendung in dieser Arbeit den Vorteil, dass die Anforderung einer hypothesenfreien Datenanalyse einfach einzuhalten ist, da ohne Bezeichnungen auch keine kausalen Zusammenhänge zwischen einzelnen Attributen ausgemacht werden können. Bezüglich fehlender Daten ist zu erwähnen, dass abgesehen vom Klassenattribut nur ein weiteres Attribut für alle Objekte eine Merkmalsausprägung aufweist. Folglich sind 169 Attribute unvollständig, wodurch nur 569 der 60.000 Objekte des Trainingsdatensatzes durch Ausprägungen aller Attribute beschrieben werden. Abseits des Klassenattributs ist der Trainingsdatensatz durch eine Ausfallrate von 7,75% gekennzeichnet.

Ferner ist der Trainingsdatensatz dadurch gekennzeichnet, dass die Werte des Klassenattributs ein starkes Ungleichgewicht aufweisen. Nur 1000 Objekte, sprich 1,67% der Fahrzeuge, gehören der Klasse *positiv* mit tatsächlich fehlerhaften APS an. Insgesamt stehen also zwei multivariate, stark unausgewogene Datensätze zur Verfügung, die zudem durch fehlende Daten in nahezu allen Merkmalen gekennzeichnet sind.

Aufgabe der Datenanalyse bzw. des KDD-Prozesses ist es, anhand der Merkmalsausprägungen fehlerhafte APS zu erkennen. Hintergrund dessen ist, dass eine korrekte Erkennung fehlerhafter APS nicht nur sicherheitsrelevant ist, sondern dem Unternehmen gleichzeitig zum Einsparen von Kosten verhelfen soll. Daher wird die Leistungsfähigkeit des KDD-Prozesses in der Aufgabenstellung des Unternehmens anhand verursachter Kosten durch Fehlklassifikation bewertet. Aus einer Einteilung in zwei Klassen folgen auch zwei potenzielle Fehlertypen, die in diesem konkreten Beispiel jeweils unterschiedliche Kosten verursachen. Bei einer falschen Zuordnung zur Klasse fehlerhafter Systeme (falsch positiv) entstehen geringe Kosten durch nicht notwendige Überprüfungen in Werkstätten. Unentdeckte Fehler im APS (falsch negativ) sind im Gegensatz dazu mit dramatischeren Folgen und dementsprechend mit höher bezifferten Kosten verbunden. Ziel der Datenanalyse ist aus Unternehmenssicht also die Kostenminimierung durch die richtige Einteilung in zwei Klassen. Die daraus abgeleitete Aufgabe für das Data Mining besteht also in der Klassifizierung. Bei der Klassifizierung auf Grundlage von Daten handelt es sich auf dem Gebiet des Data Mining um ein Vorhersageproblem, das sich mithilfe verschiedener Data-Mining-Algorithmen lösen lässt.

4.2.2 Datenvorverarbeitung

Die Datenvorverarbeitung bzw. die darin enthaltene Aufgabe des Umgangs mit fehlenden Werten stellen den Hauptteil des hier dargelegten Vorgehens dar. Daher sind die einzelnen Aufgaben dieser Phase nach CRISP-DM bereits im Kapitel über Vorgehensmodelle näher erläutert worden. Wie dort dargelegt, folgen aus der Phase der Datenvorverarbeitung ein finaler Datensatz und eine dazugehörige Beschreibung sowie ein Berichtes über die Vorverarbeitung. Da die Phase der Datenvorverarbeitung bzw. der Schritt der Imputation in dieser exemplarischen Anwendung wiederholt und unter Einfluss verschiedener Faktoren durchgeführt werden, werden in dieser Studie gleich mehrere finale Datensätze vorbereitet. Daher werden auch verschiedene, parallele Prozesse zur Erzeugung dieser Datensätze beschrieben. Außerdem ist anzumerken, dass die Aufgaben nicht in einer strikten Reihenfolge ablaufen, sondern so angeordnet werden, dass schlussendlich die finalen Datensätze für den weiteren Prozess entstehen. Dabei können einzelne Aktivitäten im Zuge der Vorverarbeitung auch wiederholt notwendig sein.

Datenauswahl

Die Datenauswahl umfasst nach CRISP-DM nicht die Auswahl des Anwendungsbeispiels, sondern die Auswahl der zu verwendenden Daten aus der zur Verfügung stehenden Datenmenge. Da die Daten für dieses Anwendungsbeispiel bereits vorselektiert vorliegen, bedarf es an dieser Stelle nicht unbedingt einer näheren Auswahl der Daten. Allerdings ist hier zu bemerken, dass die exemplarische Anwendung anhand verschiedener Stichproben durchgeführt wird, um den Ablauf sowohl anhand eines per se unvollständigen als auch mit eines vollständigen Datensatzes durchzuführen und auszuwerten. Daher wird zum einen der gesamte, unvollständige Trainingsdatensatz für die Anwendung und Auswertung verwendet. Zum anderen wird der Ablauf mit einer Stichprobe aus den vollständigen Objekten aus Trainings- und Testdatensatz durchgeführt, um die Ergebnisse des per se unvollständigen Datensatzes gleichzeitig zu validieren und außerdem den Einfluss verschiedener Ausfallraten testen zu können.

Abbildung 4-1 zeigt, wie die Stichprobe in RapidMiner erzeugt wird. Zunächst werden der Test- und Trainingsdatensatz zusammengeführt, damit eine möglichst große Anzahl an vollständigen Objekten zur Verfügung steht. Anschließend werden dann alle unvollständigen Objekte des zusammengeführten Datensatzes entfernt, bevor eine zufällige Stichprobe mit $n = 200$ Objekten unter der Bedingung gezogen wird, dass beide Klassen gleichermaßen vertreten sind.

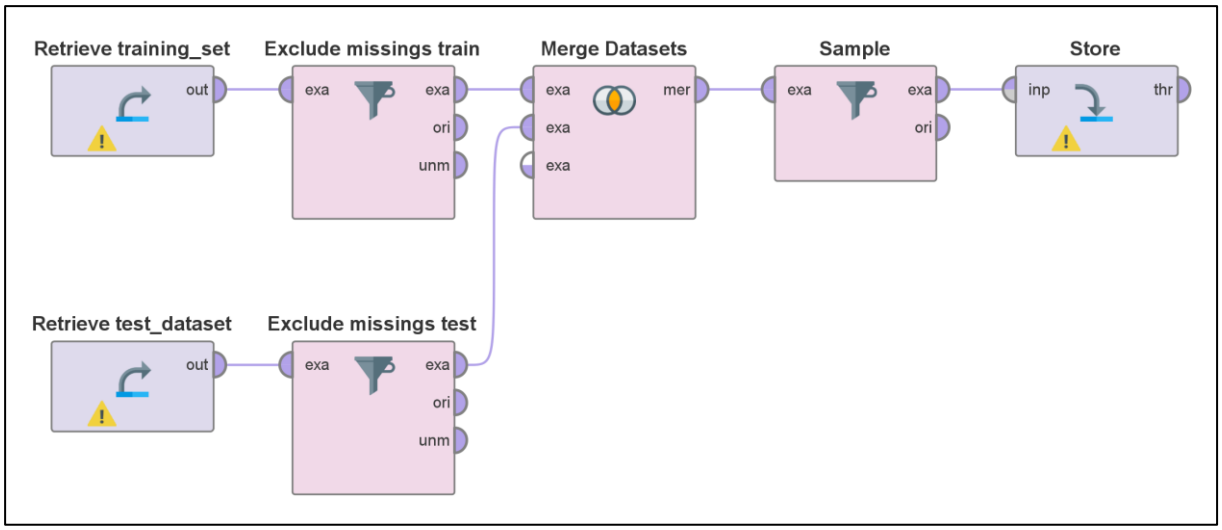

Abbildung 4-1: Erzeugung der Stichprobe für weitere Anwendung in RapidMiner

Die Größe der Stichprobe ergibt sich dabei aus der Menge der vollständigen Objekte der Klassen. Denn auch nach Zusammenführung der beiden Datensätze gehören nur knapp über 100 der vollständigen Objekte der positiven Klasse an. Durch diesen Prozess wird schließlich ein ausbalancierter, vollständiger Datensatz erzeugt, aus dem im Folgenden zunächst Werte entfernt werden, bevor diese mithilfe der ausgewählten Imputationsverfahren wieder eingesetzt werden.

Datenkonstruktion, -integration und -formatierung

Die Aufgaben der Datenkonstruktion, -integration und -formatierung können für diese Studie größtenteils vernachlässigt werden, da nicht wesentlich in die Datensätze eingegriffen wird. Es werden lediglich an gegebener Stelle Identifikationszahlen (IDs) generiert, um Objekte an mehreren Punkten innerhalb des Ablaufs eindeutig zu identifizieren und somit miteinander vergleichen zu können.

Außerdem wird in dieser Arbeit das Entfernen von Werten aus der zuvor gebildeten Stichprobe an dieser Stelle eingeordnet. Zwar stellt das Entfernen von Werten keine klassische Aktivität der Datenkonstruktion, -integration oder -formatierung dar, allerdings folgt diese Aktivität im Kontext dieser Studie zwangsläufig auf die Datenauswahl und muss im Vorfeld der Datenbereinigung geschehen. Während der per se unvollständige Trainingsdatensatz also an dieser Stelle unberührt bleibt, werden aus der zuvor erzeugten Stichprobe 10, 20, 30, 40 und 50 Prozent der Werte gemäß dem MCAR-Ausfallmechanismus entfernt.

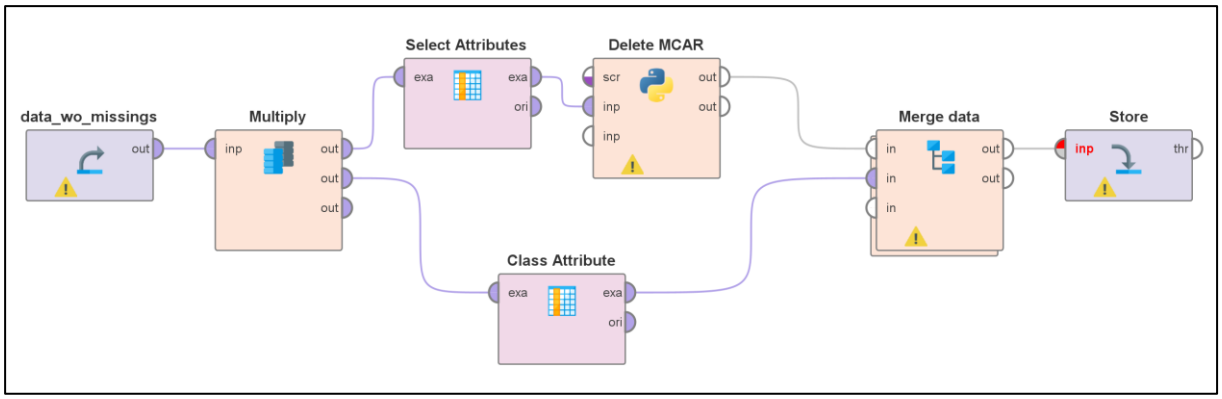

Abbildung 4-2: Vorbereiten der Stichprobe durch Entfernen von Werten in RapidMiner

Abbildung 4-2 verdeutlicht, wie das Ziehen der Stichprobe innerhalb der verwendeten Software umgesetzt wird. Um keine Werte aus des Klassenattributs zu entfernen, werden alle anderen Merkmale zunächst separiert, bevor Werte ebendieser Merkmale mithilfe eines einfachen Python-Programmbausteins gemäß den oben definierten Ausfallraten entfernt werden. Anschließend werden die Klassen wieder den jeweiligen Objekten zugeordnet, bevor die vorbereiteten Stichproben als Datensätze für den weiteren Ablauf abgespeichert werden.

Datenbereinigung und Imputation

Die Datenbereinigung umfasst nach CRISP-DM alle weiteren Schritte zur weiteren Vorverarbeitung der Daten und damit auch den Umgang mit fehlenden Merkmalswerten. Da die Datensätze an dieser Stelle abgesehen von der Imputation vollständig für den weiteren Wissensentdeckungsprozess vorbereitet sind, widmet sich dieser Abschnitt ausschließlich den folglich verwendeten Imputationsverfahren. Dabei wird zunächst kurz auf die Auswahl der Verfahren eingegangen, bevor die Implementierung in RapidMiner ausgeführt wird.

Nachdem auf Grundlage des ausgewählten Fallbeispiels nun verschiedene, unvollständige Datensätze für den weiteren Wissensentdeckungsprozess vorliegen, werden diese schließlich für die mithilfe von Imputationsverfahren als Vorbereitung für das anschließende Data Mining vervollständigt. Die Auswahl der verwendeten bzw. untersuchten Imputationsverfahren orientiert sich dabei sowohl an der systematischen Literaturrecherche als auch an der in der Software enthaltenen Algorithmen. Weitergehend erfordert die Vervollständigung des gesamten Trainingsdatensatzes deutlich mehr Rechenkapazitäten, weswegen der Großteil der Verfahren ausschließlich auf den Stichproben mit verschiedenen Ausfallraten angewendet wird. Daher werden anhand des ganzen Trainingsdatensatz die Mittelwertsimputation als traditionelles Verfahren und eine standardmäßige Implementation der multiplen Imputation respektive des FCS-Ansatzes als fortgeschrittene Methode miteinander verglichen. Auf die Stichprobe mit 200 Objekten werden zusätzlich die kNN-Imputation und zwei Entscheidungsbaummodelle zur

Imputation angewendet. Die Mittelwerts- und kNN-Imputation stellen in diesem Vergleich stellvertretend die traditionellen Methoden dar, die nach der systematischen Literaturrecherche zumindest hinsichtlich der reinen Imputationsqualität schlechtere Ergebnisse als die modernen, modellbasierten Methoden versprechen. Die Entscheidungsbaumverfahren versprechen hingegen gemäß der systematischen Literaturrecherche auch ohne aufwendige Anpassung konstant gute Ergebnisse. Die explizite Anpassung von multiplen Imputationsverfahren könnte zwar möglicherweise zu den besten Ergebnissen führen, allerdings würde die ausführliche Anpassung dieses Verfahren das Ausmaß dieser exemplarischen Anwendung übersteigen. Zudem wird in diesem Anwendungsbeispiel nicht das Ziel verfolgt, optimale Imputationsergebnisse zu erhalten, sondern vielmehr den Einfluss der Verfahren auf Wissensentdeckungsprozesse zu untersuchen. Daher wird eine standardmäßige Implementation dieses Verfahren anhand der Datensätze angewendet. Außerdem ist insbesondere für die modellbasierten Imputationsverfahren anzumerken, dass die Imputation in RapidMiner nur auf Grundlage der beschreibenden Attribute und eben nicht auf Grundlage des Klassenattributs durchgeführt wird. Der Einbezug des Klassenattributs hätte zur Folge, dass dieses Attribut zur Vorhersage von Werten genutzt würde, anhand derer anschließend wiederum die Klasse vorhergesagt würde.

Bei der Mittelwertsimputation werden die arithmetischen Mittelwerte eines Attributs für fehlende Ausprägungen ebendieses Attributs gemäß den Ausführungen im Grundlagenteil eingesetzt. Da es sich bei der Mittelwertsimputation um die standardmäßige Methode zur Imputation in RapidMiner handelt, existiert in der Software ein vorgefertigter Programmbaustein zur Umsetzung dieses Imputationsverfahrens.

Die kNN-Imputation wird ebenfalls mithilfe eines vorgefertigten Programmbausteins für die modellbasierte Imputation implementiert, indem zunächst die fünf ($k = 5$) nächsten Nachbarn eines Objektes auf Grundlage der euklidischen Distanz identifiziert werden. Im anschließenden Schritt werden die fehlenden Werte dann auf Grundlage der beobachteten Werte der identifizierten Nachbarn unter Berücksichtigung der jeweiligen Distanz abgeschätzt und eingesetzt.

Die Entscheidungsbaumverfahren zur Imputation werden innerhalb der Software ähnlich wie die kNN-Imputation implementiert. Der Programmbaustein zur modellbasierten Imputation erlernt Vorhersagemodelle anhand der vorhandenen Daten und sagt die fehlenden Werte auf dieser Grundlage vorher. Konkret werden in dieser exemplarischen Anwendung die in RapidMiner enthaltenen Modelle *Random Forest* und *Gradient Boosted Trees* verwendet, ohne die standardmäßigen Parameter dieser Modelle in der Software anzupassen.

Multiple Imputationsverfahren lassen sich nicht mit vorhandenen Bausteinen in RapidMiner umsetzen, sodass ein zusätzlicher Programmbaustein mithilfe der Programmiersprache Python erstellt zur Imputation verwendet wird. Dazu wurde die Funktion *IterativeImputer* aus der Python-Bibliothek *sklearn* als Programmbaustein in RapidMiner integriert. Diese Funktion stellt eine standardmäßige FCS-Implementation dar, die sich unter der Verwendung verschiedener Parameter auf konkrete anwendungsfälle anpassen lässt. Wie aber bereits angemerkt, wird für diese exemplarische Anwendung größtenteils von Anpassungen abgesehen. Lediglich der Parameter, der den kleinsten einzusetzenden Wert bestimmt, wird auf 0 festgelegt, um das Einsetzen negativer Werte, die in dem Datensatz sonst gar nicht vorkommen, auszuschließen. Der ansonsten standardmäßige Programmcode zur Anwendung dieser Funktion ist im Folgenden unter Algorithmus 1 angegeben und wurde darüber hinaus nur um Zeilen ergänzt, die den Datensatz aus der Software einlesen und anschließend zur weiteren Verwendung in derselben Form wieder ausgeben.

Algorithmus 1: Python-Programmcode für die multiple Imputation in RapidMiner

```
1. import pandas as pd
2. from sklearn.experimental import enable_iterative_imputer
3. from sklearn.impute import IterativeImputer
4.
5. def rm_main(data):
6.     attributes = list(data.columns)
7.
8.     imp = IterativeImputer(min_value=0)
9.     imp.fit(data)
10.
11.    df = pd.DataFrame(imp.transform(data))
12.    df.columns = attributes
13.
14.    return df
```

Die Python-Funktion verwendet das Bayesian-Ridge-Modell zur wiederholten Vorhersage und anschließenden Aggregation der Werte, weswegen der erstellte Programmbaustein in RapidMiner als FCS-Bayesian Ridge bezeichnet wird. Dementsprechend werden auch die Ergebnisse im Anschluss an dieses Verfahren im späteren Verlauf dieser Arbeit unter dieser Bezeichnung angegeben. Abbildung 4-3 veranschaulicht die Integration der multiplen Imputation

in RapidMiner. Der oben aufgeführte Programmcode ist dabei in den Operator *Execute Python* integriert. Die vor- und nachgelagerten Operatoren separieren das Klassenattribut vor der Imputation von den anderen Attributen und verhindern so, dass das Klassenattribut nicht zum Abschätzen der fehlenden Werte verwendet wird.

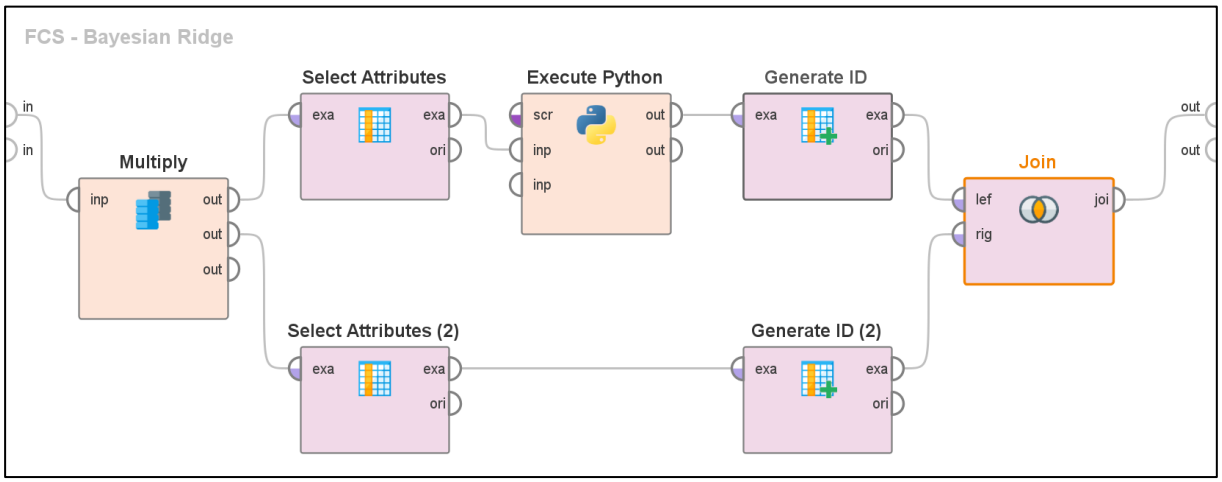

Abbildung 4-3: Programmbaustein zur multiplen Imputation in RapidMiner

4.2.3 Modellierung und Auswertung

Wie bereits in Abschnitt 4.1 dargelegt, stehen bei dieser exemplarischen Anwendung der Vergleich der Verfahren eines gesamten Wissensentdeckungsprozesses und damit die Bewertung im Anschluss an die Modellierung im Vordergrund. Um die Modellierungsfähigkeit auf Grundlage der vervollständigten Datensätze bewerten zu können, werden im Rahmen einer Kreuzvalidierung verschiedene Vorhersagemodelle anhand der Datensätze erlernt, angewendet und ausgewertet. Da es sich bei der Data-Mining-Aufgabe um die Aufgabe der Klassifizierung in zwei Klassen handelt, wird die Modellierungsfähigkeit am Anteil der korrekt eingeteilten bzw. klassifizierten Objekte gemessen.

Um zunächst geeignete Data-Mining-Algorithmen für die Anwendung und Auswertung zu identifizieren, werden diverse in RapidMiner enthaltene Algorithmen im Rahmen einer Kreuzvalidierung auf die Stichprobe mit 200 Objekten ohne fehlende Werte angewendet. Diese eingängliche Untersuchung dient dem Vergleich der Vorhersagegenauigkeit der verschiedenen Modelle anhand der tatsächlichen Daten. Dabei können besonders die entscheidungsbaumbasierten Vorhersagemodelle *Random Forest* und *Gradient Boosted Trees* eine hohe Vorhersagegenauigkeit (90%) bei gleichzeitig kleinen Abweichungen liefern. Neben diesen beiden Algorithmen wird zusätzlich ein lineares Regressionsmodell zur Vorhersage der Klassen genutzt. Dieses Vorhersagemodell kann bei dem Vergleich mit der vollständigen Stichprobe nicht mit den zuvor genannten Modellen mithalten und liefert eine vergleichsweise ungenaue Einteilung in die beiden Klassen. Dadurch werden die vervollständigten Datensätze für die Modellierung

mit zwei sehr guten und einem vergleichsweise weniger leistungsfähigen Algorithmus genutzt. Die Verwendung mehrerer Algorithmen dient in erster Linie dem Vergleich, inwiefern sich die Ergebnisse je nach verwendetem Data-Mining-Modell unterscheiden können. Die Verwendung von unterschiedlich leistungsfähigen Algorithmen kann weitergehend Aufschluss darüber geben, inwiefern die Imputationsergebnisse die Modellierungsfähigkeit beeinflussen.

Nachdem die Data-Mining-Algorithmen für die exemplarische Anwendung mithilfe des gerade dargelegten Vorgehens ausgewählt sind, werden diese Algorithmen im nächsten Schritt auf die vervollständigten Datensätze angewendet. Die Anwendung und Auswertung findet dabei mithilfe einer zehnfachen Kreuzvalidierung statt, die in RapidMiner mit einem vorgefertigten Programmbaustein umgesetzt wird. Dabei werden die Klassen der Objekte in zehn Iterationen nacheinander vorhergesagt und anschließend mit der ursprünglichen Klasse verglichen. Der jeweils betrachtete Datensatz wird bei einer zehnfachen Kreuzvalidierung in insgesamt zehn Teile unterteilt, wobei jeder dieser zehn Teile in einer Iteration als Testdatensatz fungiert, während die jeweils verbleibenden Teile als Trainingsdatensatz zum Erlernen der Modelle dienen. Dadurch wird die Klasse eines jeden Objektes im Verlauf der Kreuzvalidierung anhand der Modelle vorhergesagt und mit der originalen Klasse verglichen. Der Programmbaustein in RapidMiner wird für diese Anwendung weitergehend so angepasst, dass neben dem Datensatz mit den vorhergesagten und originalen Klassen außerdem die Anteile der korrekt Klassifizierten Objekte ausgegeben werden. Dabei werden sowohl der aggregierte Anteil der richtig eingeteilten Objekte als auch die jeweiligen Anteile für die einzelnen Klassen ausgegeben. Der Anteil der nach der Kreuzvalidierung korrekt klassifizierten Objekte wird im Folgenden als Vorhersagegenauigkeit bezeichnet und dient als hauptsächliches Auswertungskriterium für die Auswertung und Diskussion.

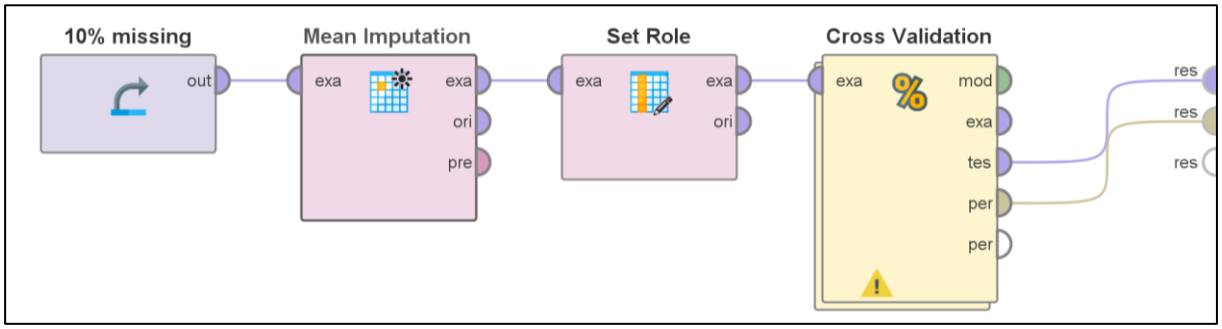

Abbildung 4-4: Ablauf bestehend aus Imputation und Kreuzvalidierung in RapidMiner

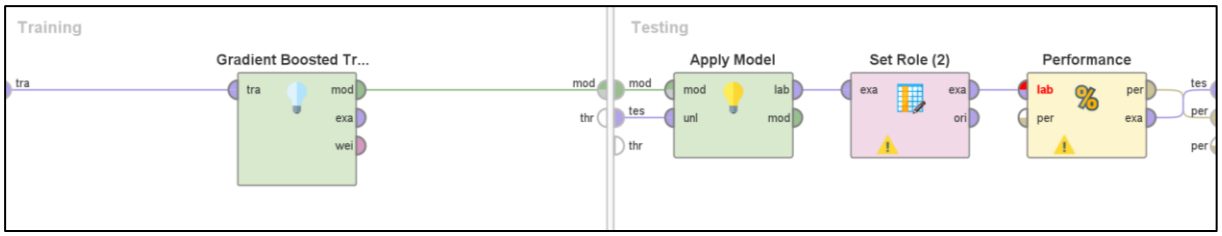

Abbildung 4-5: Kreuzvalidierungsprozess in RapidMiner

Der kombinierte Prozess aus Imputation und anschließendem Data Mining innerhalb der verwendeten Software ist in Abbildung 4-4 exemplarisch für den Ablauf mit der Stichprobe mit einer Ausfallrate von 10% und der Mittelwertsimputation dargestellt. Der zusätzliche Operator *Set Role* bestimmt dabei das vorherzusagende Zielattribut und damit in diesem Fall das Klassenattribut. Abbildung 4-5 zeigt anhand des Gradient-Boosted-Trees-Algorithmus den untergeordneten Prozess, der sich hinter dem Programmbaustein der Kreuzvalidierung verbirgt. Dieser dargestellte Ablauf wird in der Software analog auch für alle zuvor aufgeführten Datensatz- und Verfahrenskombinationen durchgeführt.

4.2.4 Ergebnisse der exemplarischen Anwendung

Nachdem in den Abschnitten 4.2.1 bis 4.2.3 ausführlich dargelegt wurde, mit welchem Vorgehen die Imputationsverfahren exemplarisch angewendet und verglichen werden, werden in diesem Abschnitt nun die Ergebnisse dieser beispielhaften Anwendung präsentiert. Die Ergebnisdarstellung dient dabei als Grundlage für die anschließende Diskussion.

Tabelle 3 führt zunächst die Vorhersagegenauigkeit nach der Verwendung verschiedener Data-Mining-Algorithmen im Anschluss an die Vervollständigung des Trainingsdatensatzes unter Zuhilfenahme der ausgewählten Imputationsverfahren auf. Die Ergebnisse sind dabei in einer Kreuztabelle respektive einer Vorhersagegenauigkeitsmatrix aufgeführt, wobei die Data-Mining-Verfahren in den Spalten und die Imputationsverfahren in den Zeilen abgebildet werden. Die Leistungsfähigkeit nach Modellierung wird dabei gemäß den vorherigen Ausführungen durch den relativen Anteil der korrekt klassifizierten Objekte beschrieben.

Tabelle 3: Matrix der Vorhersagegenauigkeiten nach Imputation und Data Mining

Verfahren	Random Forest	Gradient Boosted Trees	Lineare Regression
Mittelwertsimputation	98,66%	99,03%	99,02%
FCS-Bayesian Ridge	98,77%	99,06%	99,00%

Bei der geringen Ausfallrate von nur 7,75% und einen Datensatzgröße von 170x60.000 können alle Verfahrenskombinationen nahezu identische Ergebnisse erzielen. Die Ergebnisse sind dabei mit einer Vorhersagegenauigkeit von mindestens 98,66% allesamt auf einem sehr hohen Niveau. Ein Unterschied in der Leistungsfähigkeit des Wissensentdeckungsprozesses nach Imputation mit einem fortgeschrittenen und einem einfachen Verfahren kann demnach bei den gegebenen Bedingungen nicht ausgemacht werden. Daher wird von einer weiteren Auswertung anhand des gesamten Trainingsdatensatzes abgesehen. Im Folgenden werden somit die Ergebnisse unter Verwendung der erzeugten Stichproben ausführlicher evaluiert.

Abbildung 4-6 veranschaulicht die Vorhersagegenauigkeit nach Klassifizierung mit dem Gradient-Boosted-Trees-Algorithmus in Abhängigkeit von der Ausfallrate nach Verwendung der angegebenen Imputationsverfahren.

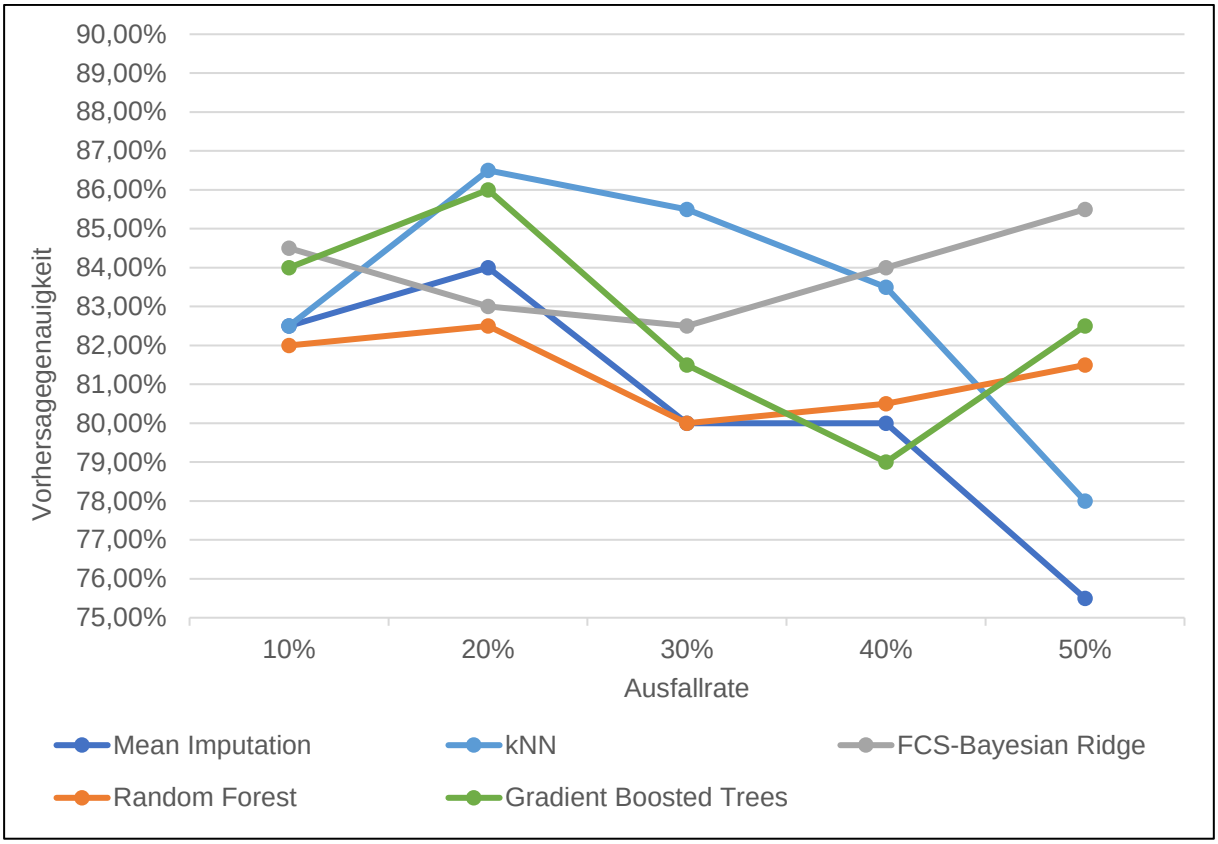

Abbildung 4-6: Vorhersagegenauigkeit nach Klassifizierung mit Gradient-Boosted-Trees-Algorithmus

Die Anteile der korrekt klassifizierten Objekte nach Vervollständigung des Datensatzes mit den verschiedenen Imputationsverfahren liegen nach dem Entfernen von 10% der Werte und anschließender Imputation im Bereich von 82% bis 84,5% und damit dicht beieinander. Bei dem Datensatz mit der Ausfallrate von 50% und ansonsten gleichen Versuchsbedingungen sind die Vorhersagegenauigkeiten deutlich weiter verstreut und liegen im Bereich zwischen 75,5% und 85,5%. Die Vorhersagegenauigkeiten nach Imputation mit den traditionellen Methoden

nimmt dabei ab einer Ausfallrate von 20% konstant ab, wohingegen die anderen Imputationsmethoden auch bei einer Ausfallrate von 50% für eine Vorhersagegenauigkeit von über 80% sorgen können. Dabei ist jedoch anzumerken, dass die Ergebnisse der Entscheidungsbaumverfahren und insbesondere die der Gradient-Boosted-Trees je nach Ausfallrate stark schwanken und dass dadurch keine Entwicklung in Abhängigkeit von der Ausfallrate auszumachen ist. Darüber hinaus sind die Vorhersagegenauigkeiten nach Imputation mit den Entscheidungsbaumverfahren bei den Ausfallraten zwischen 30% und 40% unterdurchschnittlich und können erst bei einer Ausfallrate von 50% für eine vergleichsweise gute Vorhersagegenauigkeit sorgen. Das multiple Imputationsverfahren liefert in Abhängigkeit von der Ausfallrate insgesamt die konstantesten und gerade bei Ausfallraten ab 40% die besten Ergebnisse.

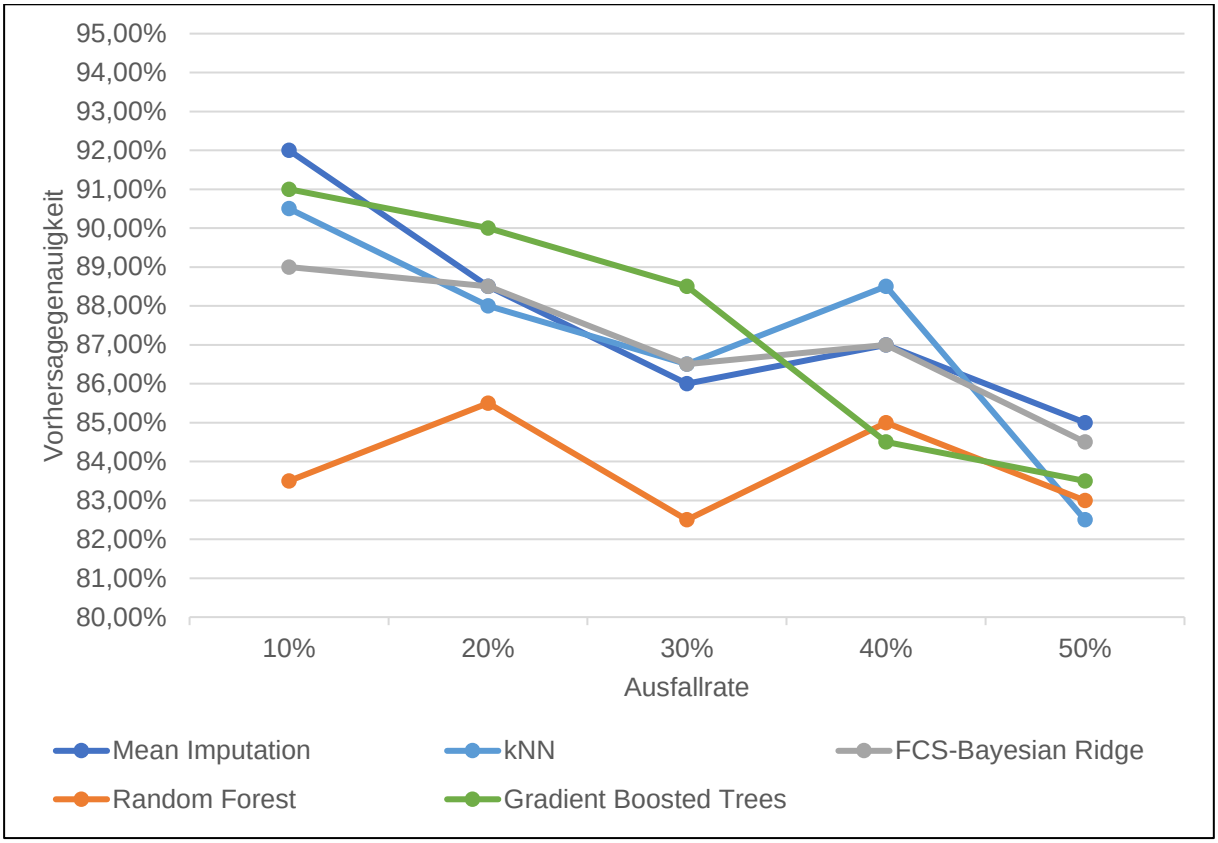

Abbildung 4-7: Vorhersagegenauigkeiten nach Klassifizierung mit Random-Forest-Algorithmus

In Abbildung 4-7 sind weitergehend die Vorhersagegenauigkeiten nach Data Mining mit dem Random-Forest-Modell abgetragen. Dieses Diagramm zeigt, dass sich die Vorhersagegenauigkeiten nach Anwendung dieser Klassifizierungsmethode insbesondere bei den niedrigeren Ausfallraten im Durchschnitt auf einem höheren Niveau befinden. Darüber hinaus lässt sich ablesen, dass die Ergebnisse nach Imputation mit allen Verfahren abgesehen von der Random-Forest-Imputation bei allen Ausfallraten dicht beieinander liegen und sich mit steigender Ausfallrate verschlechtern. Die Ergebnisse nach Random-Forest-Imputation sind im Vergleich

zu allen anderen Verfahren bei niedrigen Ausfallraten deutlich schlechter, wobei allerdings keine klare Abhängigkeit von der Ausfallrate auszumachen ist und sich die Ergebnisse damit unabhängig von der Ausfallrate auf einem konstanten Niveau bewegen. Dadurch nähern sich die Ergebnisse aller vervollständigten Datensätze mit zunehmender Ausfallrate an und liegen bei einer Ausfallrate von 50% bei Vorhersagegenauigkeiten im Intervall von 82,5% bis 85%.

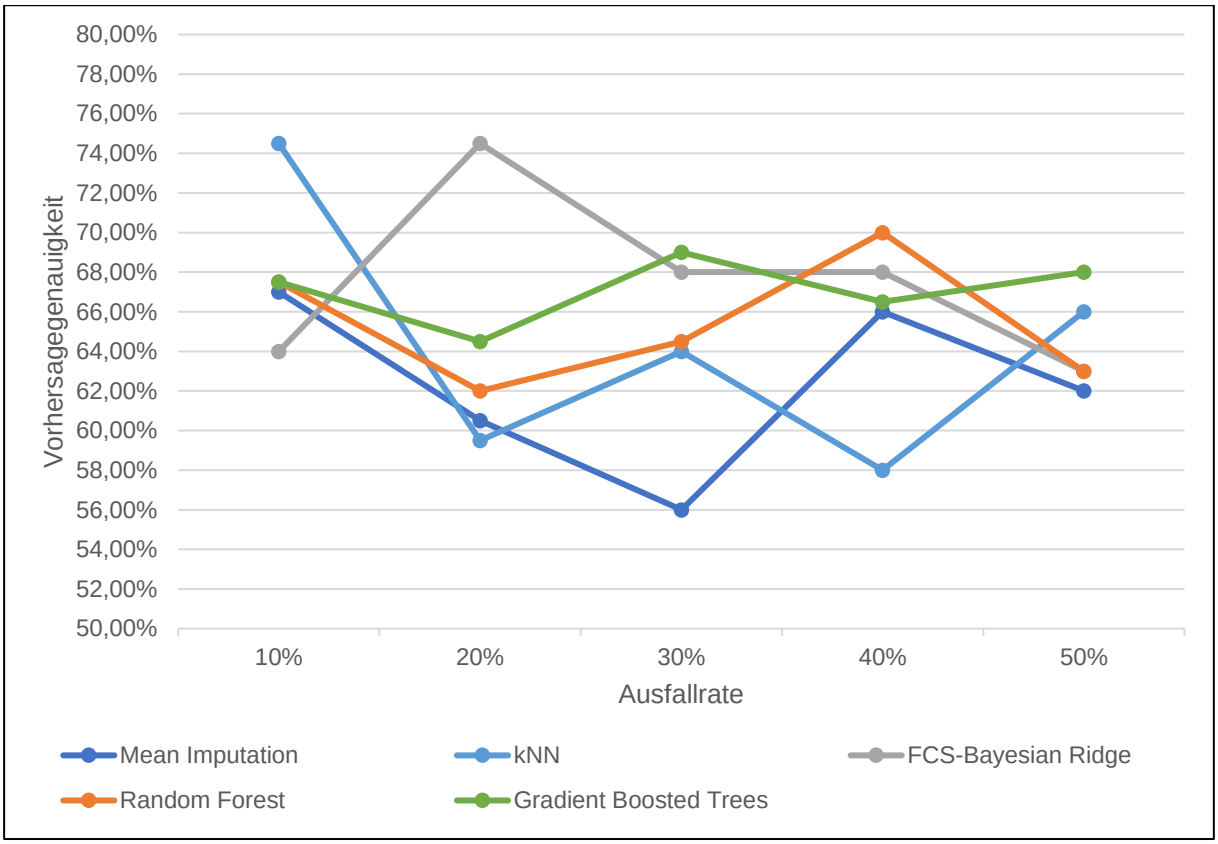

Abbildung 4-8: Vorhersagegenauigkeit nach Klassifizierung auf Grundlage eines linearen Regressionsmodells

Abbildung 4-8 veranschaulicht abschließend die Vorhersagegenauigkeiten nach Klassifizierung mithilfe des linearen Regressionsmodells und damit die Ergebnisse des zuletzt getesteten Verfahrens. Wie im vorherigen Abschnitt ausgeführt, liefert die lineare Regression gemessen an der Vorhersagegenauigkeit schon bei der vollständigen Stichprobe im Vergleich zu den zuvor getesteten Algorithmen unterdurchschnittliche Ergebnisse. Auch die Vorhersagegenauigkeiten auf Grundlage der vervollständigten Datensätze liegen mit 56% bis 75% auf einem vergleichsweise niedrigen Niveau. Aus der Zuteilung aller Objekte zu einer Klasse würde bei einem derartigen, ausbalancierten Datensatz eine Vorhersagegenauigkeit von 50% folgen. Dementsprechend stellen Vorhersagegenauigkeiten von knapp über 50% wie bei einer Ausfallrate von 30% und anschließender Mittelwertsimputation und Zuordnung durch die lineare Regression kein zufriedenstellendes Ergebnis dar. Ansonsten zeigt das Diagramm auf, dass

die Ergebnisse in Abhängigkeit von der Ausfallrate nach Imputation mit nahezu allen Verfahren stark schwanken. Lediglich bei einer Ausfallrate von 50% liegen die Vorhersagegenauigkeiten 62% bis 68% vergleichsweise dicht beieinander. Die Rangfolge der Imputationsverfahren verändert sich dabei bei jeder Veränderung der Ausfallrate.

4.3 Diskussion und Fazit

Nachdem in Kapitel 3 eine systematische Literaturrecherche und in Kapitel 4 eine exemplarische Anwendung zum Vergleich von Imputationsverfahren durchgeführt wurden, werden das Vorgehen und die Ergebnisse des vierten Kapitels diskutiert und mit den Erkenntnissen des dritten Kapitels zusammengeführt. Dabei werden die eingangs gestellten Fragestellungen zur Vergleichbarkeit noch einmal referenziert, indem ein abschließendes Fazit zur Auswahl geeigneter Imputationsverfahren auf Grundlage der beiden vorangegangen Kapitel gezogen wird.

Da die bisherige Literatur zumeist nur die reine Leistungsfähigkeit der Imputationsverfahren bewertet oder Imputations- und Data-Mining-Aufgaben vermischt, stellt die Verknüpfung zu vor- und nachgelagerten Prozessen einen bisher wenig betrachteten Ansatz dar. Die ganzheitliche Betrachtung von Imputationsverfahren im Kontext der Wissensentdeckung in Datenbanken durch die Kombination einer systematischen Literaturrecherche und einer exemplarischen Anwendung stellen ein Alleinstellungsmerkmal dieser Arbeit dar. Das Vorgehen der exemplarischen Anwendung, bei der das vorgeschlagene Rahmenkonzept in einen Wissensentdeckungsprozess nach CRISP-DM integriert wurde, konnte Erkenntnisse der systematischen Literaturrecherche zum Vorgehen derartiger Studien aufgreifen und zum Teil ergänzen. Die Software RapidMiner hat sich dabei als hilfreiches Werkzeug herausgestellt, um gesamte Wissensentdeckungsprozesse in Datenbanken abzubilden und die Ergebnisse unter Veränderung verschiedener Faktoren miteinander zu vergleichen. RapidMiner lässt sich dabei durch die Bereitstellung vorgefertigter Programmbausteine mitsamt einer jeweiligen Erklärung als besonders anwenderfreundlich und einfach verständlich beschreiben. Dadurch können auch Anwender ohne umfangreiches Fachwissen oder gar Programmierkenntnisse Wissensentdeckungsprozesse in Datenbanken durchführen und deren Ergebnisse auswerten. Programmcodes aus Programmiersprachen wie Python integrieren zu können, stellt darüber hinaus eine Möglichkeit für Anwender dar, fehlende Programmbausteine selbst anzulegen und anzupassen.

Durch die Kombination beider in dieser Arbeit angewendeten Methoden können verschiedene Ergebnisse erzeugt, miteinander verglichen und zu neuen Erkenntnissen zusammengeführt werden. Einerseits lässt die systematische Literaturrecherche die Schlussfolgerung zu, dass

moderne, modellbasierte und speziell angepasste Imputationsverfahren den traditionellen Verfahren grundsätzlich überlegen sind. Weitergehend lassen einige der betrachteten Veröffentlichungen die Annahme zu, dass die reine Imputationsqualität gemessen an der Vorhersagegenauigkeit der fehlenden Werte mit der Leistungsfähigkeit anschließender Data-Mining-Anwendungen einhergehen. Vereint man diese beiden Aussagen, so lässt sich nach der systematischen Literaturrecherche die These formulieren, dass sich die fortgeschrittenen Imputationsverfahren grundsätzlich besser für Wissensentdeckungsprozesse eignen. Die hier vorgenommene exemplarische Anwendung kann diese Annahme allerdings nicht uneingeschränkt bestätigen.

Die Ergebnisse nach Data Mining mit dem gesamten Trainingsdatensatz zeigen zunächst auf, dass die Auswahl des Imputationsverfahrens unabhängig von der Wahl des Data-Mining-Verfahrens bei einer Ausfallrate unter 10% einen eher geringen Einfluss auf die Leistungsfähigkeit des Wissensentdeckungsprozesses haben. Damit kann die exemplarische Anwendung die Ergebnisse der systematischen Literaturrecherche bezüglich niedriger Ausfallraten bestätigen. Die Ergebnisse der exemplarischen Anwendung legen darüber hinaus nahe, dass nicht nur die Imputationsqualität an sich, sondern auch die Modellierungsfähigkeit nach Imputation bei niedrigen Ausfallraten nicht ausschlaggebend von der Auswahl des Imputationsverfahrens abhängen.

Nach der exemplarischen Anwendung auf Grundlage der erzeugten Stichproben mit verschiedenen Ausfallraten können weitergehend deutlich mehr Erkenntnisse abgeleitet werden. Zunächst einmal zeigen die Abbildungen 4-6 bis 4-8 eindeutig, dass verschiedene Data-Mining-Algorithmen auf verschiedene Art und Weise auf die Veränderung der Ausfallrate und des Imputationsverfahrens reagieren. Während die Ergebnisse nach Klassifizierung mit dem Gradient-Boosted-Trees-Algorithmus mit steigender Ausfallrate größere Unterschiede aufweisen, nähern sich die Ergebnisse nach Klassifizierung mit den anderen beiden Verfahren bei höherer Ausfallrate an. Außerdem belegen die Imputationsverfahren je nach Klassifizierungsverfahren und Ausfallrate unterschiedliche Platzierungen in der Rangfolge, ohne dass ein eindeutiger Unterschied zwischen traditionellen und fortgeschrittenen Methoden auszumachen ist. Dieselben Imputationsergebnisse können also bei Veränderung des Data-Mining-Verfahrens Unterschiede in der Modellierungsfähigkeit zur Folge haben. Damit bestätigt die exemplarische Anwendung die systematische Literaturrecherche hinsichtlich der Aussage, dass sich die Qualität von Imputationsverfahren nur schwierig bzw. gar nicht allgemein bewerten lassen.

Aus den vorangegangenen Ausführungen lässt sich somit ableiten, dass Imputationsverfahren nicht nur in Abhängigkeit von bestimmten Charakteristika des konkreten Anwendungsfalls, sondern auch in Abhängigkeit vom Data-Mining-Verfahren ausgewählt werden sollten. Damit

unterstützen die Ergebnisse der exemplarischen Anwendung den Gedanken aus der Diskussion der systematischen Literaturrecherche, dass Imputationsverfahren und Data Mining für den Vergleich im Kontext von Wissensentdeckungsprozessen in Datenbanken nicht klar voneinander getrennt werden sollten. Eine losgelöste Betrachtung und Auswahl von Imputationsverfahren sind daher nicht zu empfehlen. Bei all den gerade dargelegten Erkenntnissen ist zwar zu beachten, dass es sich um eine beispielhafte Anwendung handelt, deren Ergebnisse nicht uneingeschränkt auf andere Fälle übertragen werden dürfen. Allerdings ist anzunehmen, dass die reine Imputationsqualität nicht unbedingt mit der Modellierungsfähigkeit einhergeht. Denn sowohl einzelne Studien der systematischen Literaturrecherche als auch diese exemplarische Anwendung lassen die These zu, dass sich in der Imputationsqualität unterschiedliche Datensätze unterschiedliche Modellierungsfähigkeiten in Abhängigkeit anderer Faktoren hervorrufen können.

Abschließend lässt sich festhalten, dass sowohl aus der systematischen Literaturrecherche als auch aus der exemplarischen Anwendung folgt, dass die Auswahl eines geeigneten Imputationsverfahrens eine stark einzelfallabhängige Entscheidung ist. Im Kontext von Wissensentdeckungsprozessen in Datenbanken sollte diese Entscheidung außerdem in Verbindung mit der Auswahl des Data-Mining-Verfahrens getroffen werden. Dabei ist die Auswahl eines geeigneten Imputationsverfahrens zweifelsohne von großer Bedeutung, da die Modellierungsfähigkeit eindeutig von den Imputationsergebnissen abhängig ist. Inwiefern die Imputationsergebnisse die Modellierungsfähigkeit genau beeinflussen, kann allerdings weder durch die exemplarische Anwendung noch durch die systematische Literaturrecherche geklärt werden. Vielmehr zeigen die beiden Methoden eine generelle Abhängigkeit auf, wobei das erarbeitete Rahmenkonzept und die genaue Beschreibung der exemplarischen Anwendung als Vorlage zur Auswahl eines geeigneten Imputationsverfahren dienen können.

5 Zusammenfassung und Ausblick

Dieses Kapitel fasst schließlich die Ergebnisse dieser Arbeit und explizit die Erkenntnisse der Kapitel drei und vier zusammen. Nachdem die ersten Abschnitte dieser Arbeit in relevante Bereiche der Datenwissenschaft, namentlich die Wissensentdeckung in Datenbanken und Data Mining, einführen, richtet sich der Fokus in den darauffolgenden Abschnitten zunehmend auf den Umgang mit fehlenden Merkmalswerten und insbesondere auf Imputationsverfahren.

Um Imputationsverfahren im Kontext der Wissensentdeckung in Datenbanken als Vorbereitung für Data Mining zu vergleichen, wird im dritten Kapitel zunächst eine systematische Literaturrecherche angestellt. Diese Literaturrecherche lässt auf oberster Ebene den Schluss zu, dass sich verschiedene Imputationsverfahren und deren Ergebnisse durchaus in der Qualität unterscheiden. Die Qualität der Imputation ist dabei allerdings von diversen Einflussfaktoren und Rahmenbedingungen abhängig, weswegen die Auswahl des optimalen Imputationsverfahrens in jedem Fall eine anwendungsfallabhängige Entscheidung ist. Insbesondere die differenzierte Betrachtung von Anwendungsfällen mit longitudinalen bzw. nicht-longitudinalen Datensätzen ist bei der Auswahl eines geeigneten Imputationsverfahrens von Bedeutung. Dadurch lassen sich nur eingeschränkt Schlussfolgerungen in Bezug auf die allgemeine Qualität der Imputationsverfahren ziehen. Allerdings resultiert aus einem Großteil der Studien, dass moderne, modellbasierte bzw. speziell auf den Anwendungsfall angepasste Verfahren bessere Imputationsergebnisse als einfache, traditionelle Verfahren liefern. Die qualitativen Unterschiede sind dabei maßgeblich von der Ausfallrate des untersuchten Datensatzes abhängig, wobei die Unterschiede in der Regel mit höherer Ausfallrate wachsen. Bezüglich anderer Einflussfaktoren wie dem Ausfallmuster oder -mechanismus lassen sich keine allgemeingültigen Schlussfolgerungen ableiten. Darüber hinaus lässt die systematischen Literaturrecherche keine eindeutigen Schlussfolgerungen zu der Eignung einzelner Imputationsmethoden in Bezug auf bestimmte Data-Mining-Anwendungen zu. Die Annahme, dass die Qualität der Imputationsverfahren und -ergebnisse direkt mit der Modellierungsfähigkeit korreliert und damit entscheidend für die Leistungsfähigkeit von Wissensentdeckungsprozessen in Datenbanken ist, kann aufgrund zu weniger und teilweise widersprüchlicher Resultate nicht abschließend bestätigt werden.

An diesem Punkt knüpft im Anschluss dieser Arbeit die exemplarische Anwendung verschiedener Imputationsverfahren im Kontext von Wissensentdeckungsprozessen in Datenbanken an. Die exemplarische Anwendung veranschaulicht nicht nur sehr detailliert, wie sich der Vergleich von Imputationsverfahren in Wissensentdeckungsprozesse integrieren lässt, sondern untersucht die Abhängigkeiten zwischen Imputations- und Data-Mining-Verfahren anhand ei-

nes konkreten, realen Fallbeispiels. Die Ergebnisse dieser exemplarischen Anwendung verdeutlichen, dass die Leistungsfähigkeit von Data-Mining-Anwendungen und damit die von Wissensentdeckungsprozessen in Datenbanken maßgeblich von der Auswahl der Imputationsverfahren abhängen. Gerade bei höheren Ausfallraten und kleineren Stichprobengrößen unterscheidet sich die Vorhersagegenauigkeit von Data-Mining-Algorithmen in Abhängigkeit vom gewählten Imputationsverfahren. Dabei lassen sich anhand des hier verwendeten Fallbeispiels allerdings keine allgemeinen Muster ableiten. Zwei identische, durch Imputationsverfahren vervollständigte Datensätze können demnach bei Verwendung unterschiedlicher Data-Mining-Algorithmen verschiedene Platzierungen in der Rangfolge der Leistungsfähigkeit belegen.

Die systematische Literaturrecherche und die exemplarische Anwendung verdeutlichen also die Bedeutung im Kontext von Wissensentdeckungsprozessen in Datenbanken, ohne die konkrete Eignung einzelner Verfahren für anschließendes Data Mining bewerten zu können. Die Ergebnisse der systematischen Literaturrecherche und die detaillierte Beschreibung der exemplarischen Anwendung können dabei als entscheidende Hilfe zum anwendungsspezifischen Vergleich von Imputationsverfahren im Kontext von Wissensentdeckungsprozessen in Datenbanken dienen.

Diese zuvor dargestellten Erkenntnisse bezüglich des Zusammenspiels aus Imputationsverfahren und Data-Mining-Verfahren werfen die Frage auf, ob die klare Trennung von Datenvorverarbeitung bzw. Imputation und Data Mining überhaupt sinnvoll ist. Denn auch wenn Imputationsverfahren für spezielle Anwendungsfälle optimiert werden und dadurch die qualitativ besten Imputationsergebnisse liefern, ist eine optimale Eignung im Kontext der Wissensentdeckung nicht sichergestellt. Außerdem ähneln die Abläufe zur Umsetzung von Imputationsverfahren insbesondere bei modellbasierten Verfahren sehr stark den Abläufen gesamter Wissensentdeckungsprozesse in Datenbanken, wodurch die Grenzen zwischen Datenvorverarbeitung und Data Mining ohnehin zunehmend verschwimmen.

Für zukünftige Untersuchungen sollte es also von besonderem Interesse sein, Imputationsverfahren insbesondere in Kombination mit Data-Mining-Anwendungen zu untersuchen. Dahingehend könnte eine Untersuchung zum Zusammenhang zwischen verschiedenen Auswertungsmetriken ein Anstoß für zukünftige Studien sein. Der Zusammenhang von Statistischem Verhalten, Vorhersagegenauigkeit und Modellierungsfähigkeit könnte neue Erkenntnisse hinsichtlich der Eignung von Imputationsverfahren im Rahmen von Wissensentdeckungsprozessen in Datenbanken liefern und damit bei der Auswahl von Imputationsverfahren behilflich sein.

Literaturverzeichnis

Asamoah, D. A., & Sharda, R. (2019). CRISP-eSNeP: Towards a data-driven knowledge discovery process for electronic social networks. *Journal of Decision Systems* (Vol. 28(4), pp. 286-308).

Bankhofer, Udo (1995). Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse. Doktorarbeit.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of school psychology* (Vol. 48(1), pp. 5-37).

Brachman, R. J. & Anand, T. (1996). The process of knowledge discovery in databases. In Fayyad, U. et al. (Hrsg.): *Advances in knowledge discovery and data mining* (pp. 37-58). AAAI Press / The MIT Press.

Chapman, P., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (1999). CRISP-DM 1.0. Step-by-step data mining guide. *SPSS inc* (Vol. 9(13), pp. 1-73).

Chen, P. P. S. (1976). The entity-relationship model — toward a unified view of data. *ACM transactions on database systems (TODS)* (Vol. 1(1), pp. 9-36).

Crespo Turrado, C., Sánchez Lasheras, F., Calvo-Rollé, J. L., Piñón-Pazos, A. J., & de Cos Juez, F. J. (2015). A new missing data imputation algorithm applied to electrical data loggers. *Sensors* (Vol. 15(12), pp. 31069-31082).

Donders, A. R. T., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology* (Vol. 59(10), pp. 1087-1091).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine* (Vol. 17(3), pp. 37-54).

Fayyad, U. (2005). Knowledge discovery in databases: An overview. In *Inductive Logic Programming: 7th International Workshop, ILP-97 Prague, Czech Republic September 17–20, 1997 Proceedings* (pp. 1-16). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fink, A. (2005). Conducting research literature reviews: From the Internet to paper. (2. Auflage) Thousand Oaks, CA: Sage.

Gheware, S. D., Kejkar, A. S., & Tondare, S. M. (2014). Data mining: Task, tools, techniques and applications. *International Journal of Advanced Research in Computer and Communication Engineering* (Vol. 3(10), pp. 8095-8098).

Grigorova, D., Tonchev, D., & Palejev, D. (2022). Comparison of Different Methods for Multiple Imputation by Chain Equation. *Large-Scale Scientific Computing: 13th International Conference, LSSC 2021, Sozopol, Bulgaria, June 7–11, 2021, Revised Selected Papers* (pp. 439-446). Cham: Springer International Publishing.

Hippner, H.; Wilde, K. D. (2001). Der Prozess des Data Mining im Marketing. In: Hippner, H.; Küsters, U.; Meyer, M. und Wilde, K. D. (Hrsg.): *Handbuch Data Mining im Marketing. Knowledge discovery in marketing databases* (S. 21-92). Braunschweig [u.a.]: Vieweg (Vieweg Gabler business computing).

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality and Quantity* (Vol. 34, pp. 331-351).

Jain, N., & Srivastava, V. (2013). Data mining techniques: a survey paper. *IJRET: International Journal of Research in Engineering and Technology* (Vol. *2*(11), pp. 2319-1163).

Jove, E., Blanco-Rodríguez, P., Casteleiro-Roca, J. L., Moreno-Arboleda, J., López-Vázquez, J. A., de Cos Juez, F. J., & Calvo-Rolle, J. L. (2018). Attempts prediction by missing data imputation in engineering degree. *International Joint Conference SOCO'17-CISIS'17-ICEUTE'17 León, Spain, September 6–8, 2017, Proceeding 12* (pp. 167-176). Springer International Publishing.

Kalaycioglu, O., Copas, A., King, M., & Omar, R. Z. (2016). A comparison of multiple-imputation methods for handling missing data in repeated measurements observational studies. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* (Vol. 197(3), pp. 683-706).

Kaya, A., & Turkoglu, I. (2021). Comparison of Clustering Performances of Missing Data Imputation Methods. *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1-6). IEEE.

Knobloch, B., Weidner, J. (2000). Eine kritische Betrachtung von Data Mining-Prozessen - Ablauf, Effizienz und Unterstützungspotenziale. In *Data Warehousing 2000: Methoden, Anwendungen, Strategien* (S. 345-365). Heidelberg: Physica/Springer.

Kurgan, L. A., Musilek, P. (2006). A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review* (Vol. 21(1), pp. 1-24).

Lang, K.M., Little, T.D. (2018). Principled Missing Data Treatments. *Prevention Science* (Vol. 19(3), pp. 284-294).

Little, R. J. A.; Rubin, D. B. (2020). Statistical Analysis with Missing Data. (3. Auflage) Hoboken: Wiley (Wiley series in probability and statistics).

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of pediatric psychology* (Vol. 39(2), pp. 151-162).

Lieber, D., Erohin, O., & Deuse, J. (2013). Wissensentdeckung im industriellen Kontext. *Zeitschrift für wirtschaftlichen Fabrikbetrieb* (Vol. 108(6), pp. 388-393).

Mertens, P., Bodendorf, F., König, W., Picot, A., Schumann, M., & Hess, T. (2017). Grundzüge der Wirtschaftsinformatik. (12. Auflage) Heidelberg, Berlin: Springer-Gabler.

Mistler, S. A., & Enders, C. K. (2017). A comparison of joint model and fully conditional specification imputation for multilevel missing data. *Journal of Educational and Behavioral Statistics* (Vol. 42(4), pp. 432-466).

MIT Critical Data (2016). Secondary analysis of electronic health records. Springer Nature.

Nordholt, E.S. (1998). Imputation: Methods, Simulation Experiments and Practical Examples. *International Statistical Review* (Vol. 66, pp. 157-180).

North, K. (2021). Wissensorientierte Unternehmensführung. Wertschöpfung durch Wissen. (7. Auflage) Wiesbaden: Springer-Gabler.

Okoli, C. (2015). A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems* (Vol. 37, pp. 879-910).

Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation* (Vol. 7(4), pp. 353-383).

Plaue, M. (2021). Data Science: Grundlagen, Statistik und maschinelles Lernen. Springer Berlin.

Probst, G., Raub, S., Romhardt, K. (2012). Wissen managen. (7. Auflage) Wiesbaden: Springer-Gabler.

Rockel, T. (2017). Gütevergleich von Imputationsverfahren – Eine Analyse existierender Simulationsstudien. *Epidemiology* (Vol. 160(1), pp. 34-45).

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Sánchez Lasheras, F., García Nieto, P. J., García-Gonzalo, E., Argüeso Gómez, F., Rodríguez Iglesias, F. J., Suárez Sánchez, A., Rodríguez, J. D. S., Sánchez, M. L., González-Nuevo, J., Bonavera, L., Toffolatti, L., del Carmen Fernández Menéndez, S. & de Cos Juez, F. J. (2020). Missing Data Imputation for Continuous Variables Based on Multivariate Adaptive Regression Splines. In *Hybrid Artificial Intelligent Systems: 15th International Conference, HAIS 2020, Gijón, Spain, November 11-13, 2020, Proceedings* (pp. 73-85). Cham: Springer International Publishing.

Sande, I. G. (1982). Imputation in surveys: Coping with reality. *The American Statistician* (Vol. 36, pp. 145-152).

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods* (Vol. 7(2), pp. 147-177).

Smith, F. J. (2006). Data science as an academic discipline. *Data Science Journal* (Vol. 5, pp. 163-164).

Scheidler, Anne Antonia (2017). Methode zur Erschließung von Wissen aus Datenmustern in Supply-Chain-Datenbanken. Göttingen: Cuvillier Verlag (Schriftenreihe Fortschritte in der IT in Produktion und Logistik, v.1).

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software* (Vol. 45, pp. 1-67).

van Buuren, S. (2012). Flexible Imputation of Missing Data. Boca Raton: Chapman and Hall/CRC (Interdisciplinary Statistics Series).

Waller, M. A., Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics* (Vol. 2, pp. 77-84).

Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *Annals of translational medicine* (Vol. 4(1)).

Anhang

Anhang A: Literaturtabelle der Systematischen Literaturrecherche

	Titel	Autor	Jahr	Domäne
[1]	Comparison of imputation methods for missing laboratory data in medicine	Waljee et al.	2013	Medizin
[2]	Missing traffic data: comparison of imputation methods	Li, Li & Li	2014	Verkehr
[3]	Comparison of performance of data imputation methods for numeric dataset	Jadhav, Pramod & Ramanathan	2019	Datenwissenschafft
[4]	A comparison of various imputation methods for missing values in air quality data	Zainuri, Jemain & Muda	2015	Energie und Umwelt
[5]	Missing network data a comparison of different imputation methods	Krause et al.	2018	Energie und Umwelt
[6]	Comparison of missing value imputation methods in time series: the case of Turkish meteorological data	Yozgatligil et al.	2011	Energie und Umwelt
[7]	Comparison of imputation methods for missing values in air pollution data: Case study on Sydney air quality index	Wijesekara & Liyanage	2020	Energie und Umwelt
[8]	Comparison of five iterative imputation methods for multivariate classification	Liu & Brown	2012	Datenwissenschafft
[9]	Comparison of imputation methods for end-user demands in water distribution systems	Jun, Jung & Lansey	2021	Energie und Umwelt
[10]	A comparison of multiple imputation methods for missing data in longitudinal studies	Huque et al.	2018	Gesundheit und Soziales

[11]	A comparison of multiple imputation methods for recovering missing data in hydrological studies	Hamzah et al.	2021	Energie und Umwelt
[12]	A comparison of multiple imputation methods for data with missing values	Chhabra, Vashisht & Ranjan	2017	Datenwissenschafft
[13]	A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets	Gómez-Carracedo et al.	2014	Energie und Umwelt
[14]	Performance comparison of recent imputation methods for classification tasks over binary data	Ghorbani & Desmarais	2017	Datenwissenschafft
[15]	A comparison of multiple-imputation methods for missing data in repeated measurements observational studies	Kalaycioglu et al.	2016	Medizin
[16]	Comparison of different methods for univariate time series imputation in R	Moritz et al.	2015	Datenwissenschafft
[17]	Empirical Comparison of Imputation Methods for Multivariate Missing Data in Public Health	Pan & Chen	2023	Gesundheit und Soziales
[18]	A simulation comparison of imputation methods for quantitative data in the presence of multiple data patterns	Solaro et al.	2018	Datenwissenschafft
[19]	A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: A simulation study	De Silva et al.	2018	Gesundheit und Soziales
[20]	Comparison of imputation mehods for missing rate of perceived exertion data in rugby	Epp-Stobbe, Tsai & Klimstra	2022	Sport

[21]	Comparison of imputation methods for handling missing categorical data with univariate pattern	Torres Munguía	2014	Medizin
[22]	A comparison of the effects of data imputation methods on model performance	Kim et al.	2019	Energie und Umwelt
[23]	A comparison of selected parametric and non-parametric imputation methods for estimating forest biomass and basal area	Gagliasso, Hummel & Temesgen	2014	Energie und Umwelt
[24]	Multiple imputation methods for handling missing values in logitudinal studies with sampling weights: Comparison of methods implemented in Stata	De Silva et al.	2020	Gesundheit und Soziales
[25]	A comparison of multiple imputation methods for incomplete longitudinal binary data	Yamaguchi, Misumi & Maruo	2018	Medizin
[26]	Missing data in logitudinal studies: Comparison of multiple imputation methods in real clinical setting	Rosato et al.	2021	Medizin
[27]	Comparison of Selected Multiple Imputation Methods for Continous Variables-Preliminary Simulation Study Results	Misztal	2018	Datenwissenschafft
[28]	Comparison of alternative imputation methods for ordinal data	Cugnata & Salini	2017	Datenwissenschafft
[29]	A fair comparison of tree-based and parametric methods in multiple imputation by chained equations	Slade & Naylor	2020	Medizin
[30]	Comparison of regression imputation methods of baseline covariates that predict survival outcomes	Solomon, Lokhnygina & Halabi	2021	Medizin
[31]	Performance Comparison of Imputation Methods in Building Energy Data Sets	Dhungana et al.	2021	Internet of Things

[32]	Comparison of imputation methods for missing values in longitudinal data under missing completely at random (MCAR) mechanism	Anani, Asiedu & Katsekpor	2017	Datenwissenschafft
[33]	Comparison of missing value imputation methods for Turkish monthly total precipitation data	Asian et al.	2014	Energie und Umwelt
[34]	A comparison of machine learning methods for data imputation	Platias & Petasis	2020	Datenwissenschafft
[35]	A comparison of existing methods for multiple imputation in individual participant data meta-analysis	Kunkel & Kaizar	2018	Datenwissenschafft
[36]	Comparison of Different Missing-Imputation Methods for MAIAC (Multiangle Implementation of Atmospheric Correction) AOD in Estimating Daily PM2.5 Levels	Chen et al.	2020	Energie und Umwelt
[37]	Comparison of Single and MICE Imputation Methods for Missing Values: A Simulation Study.	Pauzi et al.	2021	Datenwissenschafft
[38]	A comparison of three popular methods for handling missing data: complete case analysis, inverse probability weighting, and multiple imputation	Little, Carpenter & Lee	2022	Gesundheit und Soziales
[39]	Comparison of artificial neural network (ANN) and other imputation methods in estimating missing rainfall data at Kuantan station	Norazizi & Deni	2019	Energie und Umwelt
[40]	Comparison of Different Methods for Multiple Imputation by Chain Equation	Grigorova, Tonchev & Palejev	2022	Datenwissenschafft

[41]	Comparison of interpolation, statistical, and data-driven methods for imputation of missing values in a distributed soil moisture dataset	Kornelsen & Coulibaly	2014	Energie und Umwelt
[42]	Performance Comparison of Imputation Methods for Heart Disease Prediction	Tiwaskar & Gokhale	2022	Medizin
[43]	Comparison of gaussian processes methods to linear methods for imputation of sparse physiological time series	Nickerson et al.	2018	Medizin
[44]	Comparison of Missing Data Imputation Methods using the Framingham Heart study dataset	Psychpgyios, Ilias & Askounis	2022	Medizin
[45]	Comparison of machine learning methods for clinical data imputation among a real-world lung cancer cohort	Yang et al.	2022	Medizin
[46]	Imputation analysis for time series air quality (PM10) data set: A comparison of several methods	Shaadan & Rahim	2019	Energie und Umwelt
[47]	High-dimensional imputation for the social sciences: a comparison of state-of-the-art methods	Constantini et al.	2022	Gesundheit und Soziales
[48]	Performance Comparison of Imputation Methods for Mixed Data Missing at Random with Small and Large Sample Data Set with Different Variability	Afari & Lewis	2022	Industrie
[49]	Performance comparison of some imputation methods used in missing value (s) analysis: A simulation study	Arslan et al.	2019	Datenwissenschafft
[50]	Comparison of Imputation Methods in the Survey of Income and Program Participation	McMillan	2013	Gesundheit und Soziales

[51]	A Comparison of Data Imputation Methods Utilizing Machine Learning for a New IoT System Platform	Kalay, Çinar & Sarıçiçek	2022	Internet of Things
[52]	Comparison of imputation methods for mixed data missing at random	Heidt	2019	Medizin
[53]	Comparison of Two Newly Developed Multiple Imputation Methods for MNAR Cross-Sectional Data	Liu	2020	Datenwissenschafft
[54]	A comparison of multiple imputation methods for categorical data	Akande	2015	Gesundheit und Soziales
[55]	Comparison of Missing Data Imputation Methods for Improving Detection of Obstructive Sleep Apnea	Tokle	2017	Medizin
[56]	Comparison of Methods for Filling Daily and Monthly Rainfall Missing Data: Statistical Models or Imputation of Satellite Retrievals?	Duarte, Formiga & Costa	2022	Energie und Umwelt
[57]	Comparison of Clustering Performances of Missing Data Imputation Methods	Kaya & Turkoglu	2021	Medizin
[58]	Comparison of multiple imputation methods for categorical survey items with high missing rates: Application to the family life, activity, sun, health and eating (FLASHE) study	Liu et al.	2018	Gesundheit und Soziales
[59]	Comparison of Missing Data Imputation Methods for Leaching Process Modelling	He et al.	2015	Industrie
[60]	Comparison of Imputation Methods for Univariate Time Series	Chhabra	2023	Datenwissenschafft
[61]	Comparison of Listwise Deletion and Imputation Methods for Handling a Single Missing Response Value in a Central Composite Design	Junthopas & Wongoutong	2022	Datenwissenschafft

[62]	A comparison of multiple imputation methods for bivariate hierarchical outcomes	Diaz-Ordaz et al.	2021	Datenwissenschafft
[63]	Comparison of Imputation Methods on Retrospective Breast Cancer Data in Tanzania: A Case Study of Muhimbili and Ocean Road Hospitals	Abassi, Msengwa & Akarro	2021	Medizin
[64]	Comparison of multiple imputation methods for missing data: A simulation study	Schelhaas	2021	Datenwissenschafft
[65]	Comparison of three imputation methods on a CDC COVID-19 case surveillance database	Pheysey	2022	Medizin
[66]	A comparison of multiple imputation methods for the analysis of survival data with outcome related missing covariate values	da Silva	2023	Medizin
[67]	A Comparison of Imputation Methods for Missing Risk Factor Data from Large Real-world Electronic Medical Records for Comparative Effectiveness Studies	Samanta et al.	2018	Medizin
[68]	A comparison of six methods for missing data imputation	Schmitt, Mandel & Guedj	2015	Datenwissenschafft
[69]	Application and Comparison of Imputation Methods for Missing Degradation Data	Fan, Sun & Jiang	2015	Medizin
[70]	Comparison of the Effects of Missing Data Imputation Methods in Cohort Studies of Cardiovascular Disease	Li et al.	2023	Medizin
[71]	Handling missing data for the identification of charged particles in a multilayer detector: A comparison between different imputation methods	Riggi, Riggi & Riggi	2015	Physik

[72]	Performance Comparison of Multiple Imputation Methods for Quantitative Variables for Small and Large Data with Differing Variability	Onyame	2021	Energie und Umwelt
[73]	Comparison of Alternative Imputation Methods in the National Teacher and Principal Survey	Dial et al.	2014	Gesundheit und Soziales
[74]	A comparison of imputation and prediction methods for classification of Chagas disease	Amioka	2017	Medizin
[75]	Multiple Imputation in Complex Survey Settings: A Comparison of Methods within the Health Behaviour in School-aged Children Study	Holder, McIsaac & Pickett	2015	Gesundheit und Soziales
[76]	Comparison of statistical methods for missing data imputation in MIR-radiomics	Pinedo Taquia	2020	Medizin
[77]	Comparison and Selection Criterion of Missing Imputation Methods and Quality Assessment of Monthly Rainfall in The Central Refit Valley Lakes Basin of Ethiopia	Balcha et al.	2021	Energie und Umwelt
[78]	Comparison of data imputation using the methods of Fuzzy logic, mean and autoencoder neural network	Nogueira & Munita	2021	Archäologie

Anhang B: Auswertungstabelle der systematischen Literaturrecherche

	Typ			Datenurspung			Vollständigkeit				Ausfallmechanismus						Ausfallrate			Auswertungsmetriken		
Veröffentlichung	longitudinal	nicht longitudinal	Anzahl verwendeter Datensätze	real	simuliert	real und simuliert	vollständig	unvollständig	verwendung der vollständigen Objekte	vollständige und unvollständige	nicht explizit angegeben	MCAR	MAR	MNAR	gemischt	Aufallmechan. nicht explizit angegeben	Ausfallmechan. als untersuchter Faktor?	verwendete Ausfallraten in Prozent	Ausfallrate nicht explizit angegeben	Ausfallrate als untersuchter Faktor?	statistisches Verhalten	Vorhersagegenauigkeit	Modellierungsfähigkeit	weitere Bewertungskriterien
[1]		x	2	x			x									x		10, 20, 30		x		x	x	
[2]	x		1	x			x					x	x		x			5 bis 50		x		x		
[3]		x	6	x			x									x		10, 20, 30, 40, 50		x		x		
[4]	x		1	x			x									x		5, 10, 15, 20, 25, 30		x		x		
[5]		x	1		x		x					x	x	x	x		x	10, 20, 30, 40, 50		x	x			
[6]	x		1	x					x							x		10, 20, 50		x		x		
[7]	x		1	x				x				x						5, 10, 15, 20		x	x			
[8]		x	3			x	x					x					x	10 (s); 5, 10, 15, 20, 25, 30, 35, 40 (r)		x		x		
[9]	x		1	x			x		x							x		0,25; 0,5; 10; 20; 30; 40; 50		x		x	x	
[10]	x		2			x	x			x			x						x		x			
[11]	x		2	x			x			x						x		5, 10, 15, 20, 25, 30		x		x	x	
[12]		x	1	x			x					x						ca. 20				x		
[13]		x	3	x			x	x					x					3,85; 11,95; 23,52		x	x	x		
[14]		x	14			x	x					x						5, 10, 20, 30, 40, 50		x			x	
[15]	x		1	x			x									x		20, 50		x		x		

	Typ		Datenursprung			Vollständigkeit					Ausfallmechanismus						Ausfallrate			Auswertungsmetriken			
Veröffentlichung	longitudinal	nicht longitudinal	Anzahl verwendeter Datensätze	real	simuliert	real und simuliert	vollständig	unvollständig	verwendung der vollständigen Objekte	vollständige und unvollständige	nicht explizit angegeben	MCAR	MAR	MNAR	gemischt	Aufallmechan. nicht explizit angegeben	Ausfallmechan. als untersuchter Faktor?	verwendete Ausfallraten in Prozent	Ausfallrate nicht explizit angegeben	Ausfallrate als untersuchter Faktor?	statistisches Verhalten	Vorhersagegenauigkeit	Modellierungsfähigkeit	weitere Bewertungskriterien
[16]	x		4	x			x					x						10, 30, 50, 70		x		x		
[17]		x	2	x			x						x	x			x	ca. 40				x		
[18]		x	1		x		x					x						5, 10, 20		x		x		
[19]	x		1			x	x					x	x				x	5, 25, 50		x		x		
[20]		x	1	x			x						x					5, 10, 15, 20, 25, 30		x		x	x	
[21]		x	1	x			x					x	x	x			x	5, 15		x		x		
[22]	x		1	x			x									x		10, 20, 30, 70		x		x	x	
[23]		x	1 (k)	x			x									x			x			x	x	
[24]	x		1	x					x			x	x				x	35, 65		x		x	x	
[25]	x		1	x					x			x	x				x		x	x		x		
[26]	x		1	x					x							x		ca 35				x		
[27]		x	10	x			x					x	x	x			x	10, 20, 30, 40, 50		x		x		
[28]		x	2 (r), 1			x	x					x	x	x			x	5, 10, 20		x		x		
[29]		x	1	x				x					x					ca 50				x		
[30]		x	1	x			x						x					5, 10, 15		x		x		
[31]	x		1	x					x			x					x	5, 10, 15, 20, 25, 30, 35, 40, 45		x		x		x
[32]	x		1	x			x					x						5, 10, 15, 20, 30		x		x		
[33]	x		1 (k)	x							x		x						x			x		
[34]		x	4	x			x					x						10, 25, 50		x		x		
[35]	x	x	x			x	x						x						x			x		
[36]	x		1	x				x								x			x				x	x

	Veröffentlichung	Typ		Datenursprung			Vollständigkeit				Ausfallmechanismus						Ausfallrate			Auswertungsmetriken		
		longitudinal	nicht longitudinal	Anzahl verwendeter Datensätze	real	simuliert	real und simuliert	vollständig	unvollständig	verwendung der vollständigen Objekte	vollständige und unvollständige	nicht explizit angegeben	MCAR	MAR	MNAR	gemischt	Aufallmechan. nicht explizit angegeben	Ausfallmechan. als untersuchter Faktor?	verwendete Ausfallraten in Prozent	Ausfallrate nicht explizit angegeben	Ausfallrate als untersuchter Faktor?	statistisches Verhalten	Vorhersagegenauigkeit	Modellierungsfähigkeit	weitere Bewertungskriterien
																			5, 10, 15, 20, 25, 30, 35, 40, 45, 50						
[37]			x	1 (r), 6			x	x					x								x		x		
[38]		x		1	x				x					x						x			x	x	
[39]		x		1	x			x					x						5, 10, 15		x		x		
[40]			x	1		x		x					x						5, 10, 15		x		x		
[41]		x		1 (k)	x					x			x		x			x	ca 5, ca 20				x		
[42]			x	1	x								x						10, 20, 30, 40, 50, 60		x			x	
[43]		x		1 (k)	x					x							x			x			x		
[44]			x	1	x					x			x						10, 20, 30, 40, 50		x		x	x	
[45]		x		1	x					x			x						10, 20, 30, 40, 50		x		x	x	x
[46]		x		1	x					x							x		5, 10, 15		x		x		
[47]			x	2		x		x						x					10, 30		x		x		
[48]			x	1	x			x						x					10, 20, 30, 40, 50		x			x	
[49]			x	8		x		x									x		5, 10, 20		x		x		
[50]		x		1	x			x											ca 10			x	x		
[51]		x		1	x			x					x		x				10, 20, 30, 40		x		x		
[52]			x	2	x			x						x					10, 20, 30, 40, 50, 60		x		x		
[53]			x	2 (r), div. (s)			x	x					x	x	x			x	15, 30, 50		x		x		

	Typ			Datenursprung			Vollständigkeit				Ausfallmechanismus					Ausfallrate			Auswertungsmetriken			
Veröffentlichung	longitudinal	nicht longitudinal	Anzahl verwendeter Datensätze	real	simuliert	real und simuliert	vollständig	unvollständig	verwendung der vollständigen Objekte	vollständige und unvollständige	nicht explizit angegeben	MCAR	MAR	MNAR	gemischt	Aufallmechan. nicht explizit angegeben	Ausfallmechan. als untersuchter Faktor?	verwendete Ausfallraten in Prozent	Ausfallrate nicht explizit angegeben	Ausfallrate als untersuchter Faktor?	statistisches Verhalten	Vorhersagegenauigkeit	Modellierungsfähigkeit	weitere Bewertungskriterien
[54]		x	1	x				x				x						30, 45		x		x		
[55]		x	1	x					x			x						5, 10, 20, 30, 50		x		x	x	
[56]	x		1 (k)	x					x							x			x			x		
[57]		x	1	x			x					x						5, 15, 25, 35, 45, 60, 80		x			x	
[58]		x	1	x				x								x		ca 53			x			
[59]		x	1	x			x									x		10, 20, 30		x		x	x	
[60]	x		4	x			x									x		9,03; 19,4; 9,46; 9,26				x		
[61]		x	4		x		x									x			x		x	x		
[62]		x	1		x		x						x					20				x		
[63]		x	1 (k)	x				x				x						ca 25				x	x	
[64]		x	3		x		x						x					25, 50, 75		x		x		
[65]		x	1	x				x					x						x					
[66]		x	2			x			x				x						x			x		
[67]	x		1	x				x					x						x			x	x	
[68]		x	4	x			x					x						5, 15, 25, 35, 45		x		x	x	
[69]	x		1	x			x						x					10, 20, 30, 40, 50, 60		x		x		
[70]	x		1	x					x				x					20				x	x	
[71]		x	1		x		x								x			10, 20, 30, 40		x		x	x	
[72]	x		1			x	x					x						10, 20, 30, 40, 50		x				
[73]		x	1	x									x						x			x	x	

	Typ			Datenurspung			Vollständigkeit				Ausfallmechanismus						Ausfallrate			Auswertungsmetriken			
Veröffentlichung	longitudinal	nicht longitudinal	Anzahl verwendeter Datensätze	real	simuliert	real und simuliert	vollständig	unvollständig	verwendung der vollständigen Objekte	vollständige und unvollständige	nicht explizit angegeben	MCAR	MAR	MNAR	gemischt	Aufallmechan. nicht explizit angegeben	Ausfallmechan. als untersuchter Faktor?	verwendete Ausfallraten in Prozent	Ausfallrate nicht explizit angegeben	Ausfallrate als untersuchter Faktor?	statistisches Verhalten	Vorhersagegenauigkeit	Modellierungstähigkeit	weitere Bewertungskriterien
[74]	x		1	x					x			x		x				30 bis 40				x	x	
[75]		x	1	x				x								x		0,8 bis 23			x			
[76]		x	1	x					x							x		1, 5, 10, 20, 30, 40, 50		x		x		x
[77]	x		1	x			x									x		5, 10, 15, 20, 25		x		x		
[78]		x	1	x			x									x			x			x		