Fachlaborbericht

Datenvorverarbeitung für die Wissensentdeckung in Produktion und Logistik

Gruppe: 1

Vorgelegt von:

Jens Jakob Jacobsen (177033)

Jannik Evers (220708)

Tanja Emkrund (236397)

Melissa Hatzenbühler (210464)

Technische Universität Dortmund

Fakultät Maschinenbau

Fachgebiet IT in Produktion und Logistik

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AE	Ausreißererkennung
CRISP-DM	Cross Industry Standard Process for Data Mining
KDD	Knowledge Discovery in Databases
MAE	Mean absolute error
MAPE	Mean absolute percentage error
ML	Machine Learing
MSE	Mean square error
PDID	ParamterDescriptionId
PID	ProductID
ReS	ResultSequence
RMSE	Root Mean square error
RoS	RoutingSequence
TS	TimeStamp
WpG	WorkpieceGuid
WS	WorkSequenz
XGB	Extreme-Gradient-Boosted-Regression-Forrest

1 Einleitung

Im Rahmen der Digitalisierung wird das Erfassen von Daten weiter vorangetrieben, sei es von Nutzer- oder Prozessdaten. Freiwillig oder aufgrund gesetzlicher Bestimmungen sammelt jedes Unternehmen und jede Institution Daten. Mit zunehmender Geschwindigkeit wächst die Menge der verfügbaren Daten (Cleve und Lämmel 2020, S. 1). Die Daten werden in unterschiedlichen Bereichen eines Unternehmens verwendet, wie z. B. Geschäftsdaten, industrielle Prozessdaten, Textdaten oder auch Bild- und Videodaten (Runkler 2016, S. 1). Große, schnell wachsende Datenmengen werden auch Big Data genannt. Dabei können diese Datenmengen aus unterschiedlichsten Daten bestehen sowie strukturiert und unstrukturiert sein (Raheem 2019, S. 13). Die Bedeutung der Datenanalyse für Produktion und Logistik nimmt stetig zu, da auch durch Technologien der Industrie 4.0 zunehmend mehr Daten erzeugt werden, welche für Prozess- und Systemoptimierungen genutzt werden können (Freitag et al. 2015, S. 39). Dabei werden täglich Petabytes an Daten erzeugt und gespeichert, was zu einem enormen Datenzuwachs und Volumen an neuen Informationen führt (Luengo et al. 2020, S. 1). In der Folge steigt jährlich der prognostizierte Umsatz von Big-Data-Lösungen weltweit bis 2026 auf 92,2 Milliarden US Dollar (Statista (2023)). Entscheidend für Unternehmen ist jedoch nicht die Datenmenge, sondern ob die Daten nutzbringend verwendet werden können. Big-Data-Datenbestände können analysiert werden, um Erkenntnisse zu gewinnen, die zu besseren Entscheidungen und strategischen Geschäftsbewegungen führen (Raheem 2019, S. 13).

In der Forschung ist seit 1996 Konsens: Die Geschwindigkeit der Datenerfassung ist so hoch, dass computergestützte Werkzeuge zur wirksamen Entdeckung von verwertbarem Wissen für die Menschen von Nutzen sein werden. Die für diesen Zweck formulierten Modelle und Werkzeuge sind Gegenstand der Knowledge Discovery in Databases (KDD) (deutsch: Wissensentdeckung in Datenbanken). Der Kern des Prozesses ist die Anwendung spezifischer Data-Mining-Methoden zur Mustererkennung und -extraktion (Fayyad et al. 1996, S. 34). Die Entdeckung von Wissen in Datenbanken wird immer relevanter, da diese es Unternehmen ermöglicht, gewinnbringende Muster und Trends in ihren bestehenden Datenbeständen zu erkennen (Larose und Larose 2014a, S. 2). Aus diesem Grund werden Daten im betriebswirtschaftlichen Kontext als Ressource betrachtet, welche es ermöglicht, Wissen als entscheidende wirtschaftliche Komponente zu generieren (Plaue 2021, S. 1).

Der KDD-Prozess nach Fayyad et al. (1996) ist eines der ersten formalen Vorgehensmodelle. Der 1999 entwickelte CRISP-DM Prozess hat sich als Standard für KDD-Vorgehensmodell in der Industrie etabliert (Cleve und Lämmel 2020, S. 4). Dieses wird in Kapitel 2.1.1 näher erläutert. Ein Prozessschritt des CRISP-DM Modells sowie des KDD-Prozesses stellt die Datenvorverarbeitung dar. Die folgende Abbildung 1-1 zeigt den Aufwand der Datenvorverarbeitung im Vergleich zu den anderen Prozessschritten, basierend auf einer Studie von Kurgan und Musilek (2006).

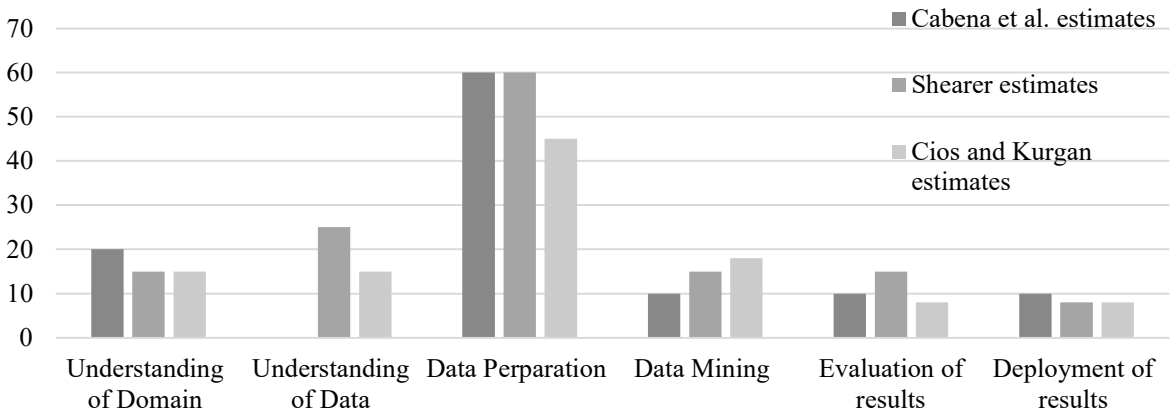

Abbildung 1-1: Aufwand je Phase (Kurgan und Musilek 2006, S. 17)

Aggarwal (2015, S. 2) stellt ebenfalls fest, dass die Datenvorverarbeitung einen großen Einfluss auf die Qualität der Datenanalyse und damit der Ergebnisse des Data Minings hat. Aufgrund des hohen Aufwands am Gesamtprojekt sowie des Einflusses auf die Qualität der Ergebnisse ist die Datenvorverarbeitung als besonders erfolgskritisch zu betrachten. Es liegt nahe, die Herausforderungen dieser Phase zu untersuchen, um die Erfolgswahrscheinlichkeit von KDD-Projekten zu verbessern.

1.1 Beschreibung der Problemstellung

Innerhalb dieser Arbeit gilt es, den Einfluss von Ausreißern in einem Datensatz auf eine Data-Mining-Fragestellung zu untersuchen und zu bestimmen, inwieweit dieser durch Datenvorverarbeitung beeinflusst werden kann. Hierfür liegt ein Datensatz vor, der genutzt wird, um einerseits eine Data-Mining-Fragestellung zu erarbeiten sowie die Ausreißererkennung zu verproben. Es handelt es sich um automatisch erzeugte Daten aus einer industriellen Fertigung. Das Unternehmen und der genaue Kontext der Daten ist für die hier bearbeitete Aufgabenstellung nicht gegeben. Schwerpunkt des vorliegenden Berichts ist folglich sowohl die Datenvorverarbeitung als auch die Erprobung verschiedener Ausreißererkennungs- und Imputationsverfahren und dessen Evaluation.

Die Datenvorverarbeitung ist ein wesentlicher Schritt bei der Analyse und Modellierung von Daten, da sie die Rohdaten in eine geeignete Form zur Identifizierung von Mustern, Trends und möglichen Beziehungen bringt. Die Hauptanforderung dieser Aufgabe ist die Identifizierung und Entfernung von Verunreinigungen durch fehlerhafte Datenobjekte, um die Reinheit der Daten für die Weiterverarbeitung zu gewährleisten (Larose und Larose 2014b, S. 17). Innerhalb der Datenvorverarbeitung ist die Erkennung von Ausreißern im Laufe der Zeit zu einer bedeutsamen Aufgabe des Data Mining geworden, da Ausreißer abweichende Muster in den Daten widerspiegeln. Aufgrund der weiten Verbreitung von Data-Mining-Techniken in verschiedenen Anwendungsbereichen ist heute eine Vielzahl von Methoden zur Erkennung von Ausreißern verfügbar, die an die jeweiligen Anforderungen angepasst werden können. Es ist daher eine Herausforderung, eine geeignete Methode zu wählen, da diese den Erkennungsprozess und das Endergebnis beeinflusst (Ranga Suri et al. 2019, S. 15)**.** Ausreißer können beispielsweise eine wesentliche Auswirkung auf die Ergebnisse eines Data-Mining-Projektes haben. Demzufolge ist es notwendig, diese Ausreißer zu erkennen und zu behandeln (Domanski 2020, S. 439).

In diesem Bericht werden verschiedene Methoden zur Ausreißerbehandlung in der Datenvorverarbeitung vorgestellt. Dazu gehören statistische Ansätze, die auf Annahmen über die Verteilung der Daten basieren. Des Weiteren werden auch distanzbasierte Ansätze wie das Clustering für die Ausreißererkennung betrachtet. Es ist wichtig zu beachten, dass die Auswahl der geeigneten Methode zur Ausreißerbehandlung von der spezifischen Analyse, dem Kontext der

Daten und den Zielen der Datenverarbeitung abhängt (Wan et al. 2019, S. 173827; Aggarwal 2017, S. 7).

1.2 Aufbau der Arbeit

Der vorliegende Bericht ist in mehrere Kapitel unterteilt. Das Hauptziel dieser Arbeit besteht in der Entwicklung einer konkreten Anwendung, um die oben genannte Fragestellung zu beantworten. Die Fragestellung wird mit Hinzunahme des Datensatzes entwickelt. Diese stellt die Grundlage dar, um neues Wissen zu gewinnen und somit einen Mehrwert aus den Daten zu generieren. Anschließend werden Ausreißerkennungs- und -behandlungsverfahren auf genau diese Fragestellung erprobt, um die Fragestellung des Fachlabors zu beantworten.

Das erste Kapitel dient als Einführung und Motivation für die Arbeit sowie zur Hervorhebung der Relevanz der Thematik. Daraufhin wird in der Problemstellung die Aufgabenstellung sowie das Ziel der Arbeit genannt, ebenso wie die Besonderheit der Datenvorverarbeitung und der Ausreißerbehandlung.

In Kapitel 2 werden die grundlegenden Konzepte des Data Mining vorgestellt. Im Rahmen dieser Darstellung wird das CRISP-DM-Modell im Detail erläutert, wobei die verschiedenen Phasen des Modells im Kontext genauer beschrieben werden. Der Fokus liegt dabei vor allem auf der Datenauswahl, Datenvorverarbeitung, Modellbildung und der Bewertung der Ergebnisse. Außerdem widmet sich Kapitel 2 dem Thema der Daten und Datenqualität. Hier werden Rahmenbedingungen festgehalten, inwieweit Daten relevant sind, um in einem KDD-Prozess untersucht zu werden. Diese Annahmen werden im weiteren Verlauf auf die vorliegenden Daten angewendet. Des Weiteren erfolgt in dem Kapitel eine strukturierte Literaturrecherche zum aktuellen Stand der Technik der Datenvorverarbeitung für die Erkennung von Ausreißern. Das verwendete Vorgehensmodell für die Recherche wird in Kapitel 2.2 beschrieben. Die Ergebnisse werden abschließend in Kapitel 2.3 analysiert und die verschiedenen Ansätze zur Ausreißererkennung vorgestellt, darunter z.B. statistische Methoden, distanzbasierte Ansätze und clusterbasierte Ansätze.

Kapitel 3 thematisiert die Entwicklung einer geeigneten Fragestellung. Zunächst werden in Abschnitt 3.1 neben der formalen Betrachtung die Daten inhaltlich untersucht, dafür werden im Weiteren die einzelnen Attribute oder Attributkombinationen erläutert. Daraufhin wird ein Data Dictionary erstellt, in welchem die Attribute dokumentiert werden. Es werden Zusammenhänge zwischen den einzelnen Attributen ermittelt, um weitere Erkenntnisse für die Entwicklung der Fragestellung zu gewinnen. Weiterhin werden die Bedingungen für den Einsatz von Ausreißerdetektierungsverfahren berücksichtigt. Durch die Eingrenzungen ist es möglich, eine Anzahl von Unterräumen des Datensatzes auszuwählen und mit diesen die Bearbeitung fortzufahren. Mit den gefilterten Daten wird zusätzlich ein ganzheitlicher Datenraum entwickelt, indem mehrere Datensätze miteinander verknüpft werden. In Abschnitt 3.2 wird ein geeignetes Modell gewählt und implementiert. Basierend auf den Anforderungen aus dem Modell werden die Daten entsprechend vorverarbeitet. Das Modell wird anschließend geeignet parametrisiert und dokumentiert. Nachdem die Datenvorverarbeitung abgeschlossen ist, werden in Abschnitt 3.3 die Ausreißererkennungsverfahren implementiert sowie die Imputationsverfahren erläutert. Damit ist die Grundlage für die Erprobung einzelner Ausreißerbehandlungsverfahren und für die Prüfung ihrer Eignung geschaffen.

Kapitel 4 thematisiert die Evaluation der Ausreißererkennungsverfahren basierend auf den Ergebnissen des Data-Mining-Modells. Es wird ein Versuchsplan vorgestellt, der die Grundlage der Erprobung der einzelnen Verfahren darstellt. Anhand ausgewählter Fehlermetriken werden unterschiedliche Verfahren miteinander verglichen. Auf der Basis der Ergebnisse werden mögliche Erkenntnisse für die Anwendung von vergleichbaren Data-Mining-Fragestellungen im industriellen Betrieb aufgestellt.

Abschließend werden im letzten Kapitel die Hauptaussagen der einzelnen Kapitel zusammengefasst und ein Fazit der Ergebnisse der Arbeit gegeben. Aus diesem Fazit wird im letzten Schritt ein Ausblick auf weitere Forschungsfragen abgeleitet.

2 Identifikation von Methoden zur Vorverarbeitung von Ausreißern

Das folgende Kapitel besteht aus zwei Teilen. Im ersten Teil (Kapitel 2.1) werden allgemeine Grundlagen zu Data-Mining-Prozessen eingeführt, die benötigt werden, um die aufgeworfene Fragestellung zu untersuchen. Hierzu zählt die Betrachtung von Vorgehensmodellen für das Data Mining sowie eine detaillierte Aufschlüsselung der zu durchlaufenden Phasen. Im zweiten Teil (Kapitel 2.2) folgt eine strukturierte Recherche zum aktuellen Stand der Technik der Datenvorverarbeitung für die Erkennung von Ausreißern. Im letzten Teil (Kapitel 2.3) werden die Ergebnisse der Recherche hinsichtlich der Fragestellung analysiert.

2.1 Data Mining Grundlagen

Wissensentdeckung aus Datenbanken (Knowledge Discovery from Databases (KDD) ist definiert als der nicht-triviale Prozess der Entdeckung von neuem und nützlichem Wissen aus Datenbanken (Kurgan und Musilek 2006, S. 2). Das Data Mining ist im KDD-Prozess der Schritt der Mustererkennung in den Daten (Mariscal et al. 2010, S. 143; Fayyad et al. 1996, S. 42). Der Begriff des Data Minings wird jedoch auch Synonym verwendet mit dem gesamten KDD-Prozess (Mariscal et al. 2010, S. 137). Data Mining ist ein komplexer Prozess, der neben dem Einsatz geeigneter Werkzeuge auch auf ein Vorgehensmodell angewiesen ist (Wirth und Hipp 2000a, S. 30). Das Verwenden eines Vorgehensmodell soll den Erfolg des Data-Mining-Vorhabens erhöhen und ein blindes Anwenden von Methoden vermeiden (Kurgan und Musilek 2006, S. 2–3). Der Erfolg von einem Data Mining Projekt ist somit abhängig von der Verwendung eines geeigneten Vorgehensmodells (Wirth und Hipp 2000a, S. 30). Es gilt folglich, ein geeignetes Vorgehensmodell zu identifizieren.

2.1.1 Data Mining Vorgehensmodell

Für dieses Projekt wird der Cross Industry Standard Process for Data Mining (CRISP-DM) als Vorgehensmodell gewählt. In Industrieprojekten hat sich das CRISP-DM Modell als De-facto-Standard etabliert (Martinez-Plumed et al. 2020, S. 2; Schröer et al. 2021, S. 533; Lieber et al. 2013, S. 389). Die weite Verbreitung und die Technologieoffenheit des Modells sind die ausschlaggebenden Kriterien für die Verwendung des Modells in diesem Projekt (Schröer et al. 2021, S. 529). Es handelt sich bei dem CRIP-DM um ein organisatorisches Modell, das nicht auf ein konkretes Werkzeug beschränkt ist (Schröer et al. 2021, S. 532).

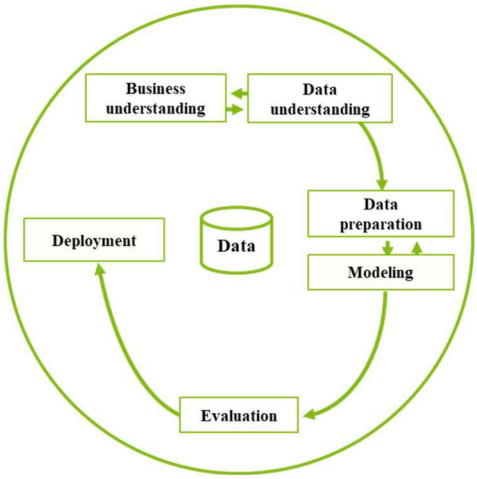

Abbildung 2-1: CRISP-DM Prozess (Wirth und Hipp 2000b, S. 33)

Das CRISP-DM Modell teilt sich mit anderen Vorgehensmodellen den iterativen Ansatz, der die einzelnen Phasen des Vorgehens mehrfach durchläuft, basierend auf den jeweiligen Phasenergebnissen (Kurgan und Musilek 2006, S. 4–5). Jeder Phase im CRISP-DM Modell werden konkrete Phasenergebnisse zugeordnet (Wirth und Hipp 2000a, S. 34). In Abbildung 2-1 ist der Ablauf des CRIP-DM Modell dargestellt. Die bidirektionalen Pfeile in der Abbildung weisen darauf hin, dass Wechselwirkungen zwischen den einzelnen Phasen erwartetet werden und es sich um einen iterativen Prozess handelt. Die sechs Phasen, aus denen das CRISP-DM Modell besteht, werden in Abschnitt 2.1.2 im Detail erläutert.

2.1.2 Der Cross Industry Standard Process for Data Mining

CRISP-DM besteht aus sechs Phasen: Business Understanding, Data Understanding, Data Preprocessing, Modeling, Evaluation und Deployment. Im folgenden Abschnitt werden diese Phasen, die Phasenergebnisse sowie weitere zugehörige Konzepte erläutert, beginnend mit dem Business Understanding.

Im Rahmen des *Business Understanding* sollen die Aufgaben und die Zielstellung des Projektes erarbeiten, sowie der Ist-Zustand erfasst werden. In diesem Rahmen werden verfügbare Ressourcen bestimmt und ein Projektplan erstellt. Hierzu zählt auch die Identifikation von Technologien und Werkzeuge, die im Projekt verwendet werden können (Saltz 2021, S. 2338; Sharma und Osei-Bryson 2008). Als Erstes werden die Zielkriterien aufgestellt (Nino et al. 2015, S. 1373). Aus dem Zielbild kann daraufhin eine Data Mining Fragestellung abgeleitet werden (Wirth und Hipp 2000a, S. 33). Die Planung und das Verständnis der Fragestellung sind essenziell für den Erfolg des Projektes (Sharma und Osei-Bryson 2008, S. 3). Am Ende der Phase steht als Phasenergebnis die Data-Mining-Fragestellung sowie der Projektplan (Sharma und Osei-Bryson 2008, S. 3).

Auf das Business Understanding folgt das *Data Understanding*. In dieser Phase werden die zu verwendenden Daten gesammelt und zusammengetragen (Wirth und Hipp 2000a, S. 33). Diese Daten werden analysiert und beschrieben, hierbei handelt sich um eine beschreibende Analyse als Metaanalyse der vorliegenden Daten. Es sollen Informationen über Syntax und Semantik der zu verwendenden Daten zusammengestellt werden (Sumana und Kweku-Muata 2010, S. 15). Die Ergebnisse der Metaanalyse werden in einem Report festgehalten (Saltz 2021, S. 2338). Neben der Syntax und Semantik der Daten wird in diesem Kontext auch die Qualität der Daten überprüft und dokumentiert. Die in diesem Rahmen identifizierten Herausforderungen der Datenqualität sind für die Phase der Datenvorverarbeitung entscheidend, da die Herausforderungen in der Datenvorverarbeitung adressiert und gelöst werden müssen (Diop et al. 2017, S. 1). Das Thema der Datenqualität wird in Abschnitt 2.1.3 gesondert untersucht. Das Phasenergebnis des Data Understandings ist ein Report, der die vorliegenden Daten beschreibt, erklärt wie diese erfasst worden sind und die Qualität der Daten festhält (Wirth und Hipp 2000b, S. 34).

Die nächste Phase, die *Datenvorverarbeitung*, ist ein entscheidender und kritischer Schritt in der Datenanalyse und -modellierung. Sie umfasst eine Vielzahl von Techniken und Verfahren, die darauf abzielen, die Datenqualität zu verbessern, die Daten für die Analyse vorzubereiten und unerwünschte Effekte oder Störungen zu reduzieren (Larose und Larose 2014b, S. 17). Innerhalb dieser Phase wird der Datensatz so gestaltet, dass er von einem Data-Mining-Algorithmus interpretiert werden kann (García et al. 2015, S. 11). Die Qualität der Daten hat einen entscheidenden Einfluss auf die Ergebnisse des Data Mining, weshalb dieser Schritt im CRISP-DM als besonders bedeutsam gilt (Nematzadeh et al. 2020, S. 17). Ein großer Anteil des Aufwandes in einem Data-Mining Projekt entfällt auf diese Phase (Saltz 2021, S. 2338; Aljaž und Mirjana Kljajić 2020, S. 79–80; Han et al. 2008, S. 96; Bilal et al. 2022, S. 107764). Abhängig vom Datensatz kann die Datenvorverarbeitung allein 10-60 % des gesamten Zeit- und Arbeitsaufwands für den Data-Mining-Prozess ausmachen (Larose und Larose 2014b, S. 17; Kurgan und Musilek 2006, S. 17).

Das Ziel der Datenvorverarbeitung besteht darin, einen finalen Datensatz zu generieren, der durch verschiedene Vorverarbeitungsschritte erreicht wird (Wirth und Hipp 2000b, S. 33–34; Sharma und Osei-Bryson 2008, S. 1–3; Sumana und Kweku-Muata 2010, S. 15–16). Folgende grundlegende Schritte können dabei die Datenvorverarbeitung, abhängig vom Datensatz, umfassen: Die Auswahl der Daten, Datenbereinigung, Konstruktion der Daten, Datenintegration und Datenformatierung (Saltz 2021, S. 2338; Chapmann et al. 2000, S. 23; García et al. 2015, S. 12; Han et al. 2008, S. 96).

Der erste Schritt der Datenvorverarbeitung umfasst die Auswahl von Daten basierend auf Kriterien wie Relevanz für die Data-Mining-Ziele, Datenqualität und technische Beschränkungen wie Datenvolumen oder Datentypen. Insbesondere bei großen Datensätzen dient dieser Schritt zur Reduzierung des Datenumfangs (García et al. 2015, S. 54; Aggarwal 2015, S. 37–38, 2015, S. 28). Hierbei werden, im Fall von strukturierten Daten, sowohl Attribute (Spalten) als auch Datensätze (Zeilen) in einer Tabelle ausgewählt (García et al. 2015, S. 54; Aggarwal 2015, S. 37–38, 2015, S. 28; Larose und Larose 2014b, S. 24). Anschließend werden die ausgewählten Daten im nächsten Schritt bereinigt.

Die Datenbereinigung bezieht sich auf das Korrigieren von Fehlern in den Daten und kann unterschiedliche Herangehensweisen beinhalten (García et al. 2015, S. 11). Fehler in den Daten können falsche Werte, fehlende Werte und Rauschen darstellen (García et al. 2015, S. 44; Aljaž und Mirjana Kljajić 2020, S. 81; Aggarwal 2015, S. 28). Methoden zur Fehlerbehebung umfassen das Löschen und Ersetzen fehlerhafter Werte (Pearson 2002, S. 62; Alexandru Prisacaru et al. 2019, S. 1; Aggarwal 2015, S. 36–37). In bestimmten Situationen können Fehler auch ignoriert werden (Chen et al. 2021, S. 341).

In den darauffolgenden Phasen werden neue Attribute, vollständig neue Datensätze oder transformierte Werte für bestehende Werte erzeugt. Informationen aus verschiedenen Tabellen oder Datensätzen werden kombiniert, um neue Datensätze oder Werte zu erstellen. Die Datenformatierung dient dazu, die Daten nach Bedarf umzuformen (Saltz 2021, S. 2338; Chapmann et al. 2000, S. 25; Larose und Larose 2014b, S. 26). Die Skalierung von Attributen in einen einheitlichen Wertebereich ermöglicht den Vergleich verschiedener Werte für Algorithmen (García et al. 2015, S. 13; Aggarwal 2015, S. 37).

Die Ergebnisse der Datenvorverarbeitung sind vollständig bearbeitete Daten. Die Datenqualität wurde verbessert, und die Daten befinden sich in einem geeigneten Zustand, der eine effektive Analyse und Verarbeitung ermöglicht (Larose und Larose 2014b, S. 17).

Die Phase *Modeling* umfasst die Auswahl und Parametrisierung geeigneter Data-Mining-Modelle (Wirth und Hipp 2000a, S. 34). Die Eignung des Modells ist im Kontext der gestellten Data-Mining-Fragestellung zu beurteilen. Im Rahmen des Machine-Learning gestützten Data Minings werden drei Kategorien von Aufgaben unterschieden: Überwachtes Lernen, unüberwachtes Lernen und bestärkendes Lernen (Ray 2019, S. 35). Die überwachten Methoden zielen darauf ab mit Hilfe von Input Daten ein Zielattribut zu prognostizieren. Die Klassifikation des Zielattributes kann weiter unterschieden werden in kategorische und kontinuierliche numerische Werte. Im Fall der kategorischen Zielvariablen handelt es sich um Klassifikation und im Fall der kontinuierlichen numerischen Zielvariablen um eine Regression (Bertolini et al. 2021, S. 3). Dem gegenüber steht das unüberwachte Lernen. Hier liegt keine Zielvariable vor, gegen die prognostiziert werden soll. Stattdessen ist das Ziel, die gegebenen Daten nach Mustern zu untersuchen. Ein Ansatz, der das illustriert ist das Clustering. Beim Clustering wird der Datensatz in Gruppen unterteilt, sodass die Datenpunkte in einer Gruppe am Ende ähnlicher zueinander sind als zu Datenpunkten in einer anderen Gruppe (Bertolini et al. 2021, S. 3). Im Anschluss an die Auswahl des Modells folgt das Erstellen der Modellinstanz und hiermit Verbunden das Durchführen von Tests zur Prüfung der technischen Validität der Data-Mining-Ergebnisse. Phasenergebnisse der Modelling Phase sind sowohl das lauffähige Modell als auch die Dokumentation der Modellentwicklung, das Vorgehen zur Parametrisierung und die Dokumentation der technischen Validierung des Modells (Chapmann et al. 2000, S. 27–29). Aufgrund der konkreten Anforderung von Data-Mining-Modellen an die Form und den Umfang

der Daten ist die Phase den Modellings in enger Wechselwirkung mit der Datenvorverarbeitung. Abhängig vom gewählten Modell, müssen die zu verwendeten Daten gegebenenfalls anders transformiert oder reduziert werden (Wirth und Hipp 2000b, S. 34). Auf die technische Evaluation des Modells folgt die Evaluation der Data Mining Ergebnisse.

In der nächsten Phase (*Evaluation*) wird das Modell evaluiert und überprüft, ob die zu Beginn festgelegten Ziele erreicht wurden. Das Ziel der Phase ist es, zu entscheiden, ob das Modell verwendbar ist, oder in einem weiteren Zyklus des CRISP-DM überarbeitet werden muss (Wirth und Hipp 2000b, S. 34).

Für die Evaluation stehen verschiedene Methoden und Performancekennzahlen zur Verfügung. Die genaue Vorgehensweise und Wahl der Methodik ist abhängig von der gewählten Machine-Learning-Aufgabe. Wie sich in Kapitel 3 herausstellt, wird für diese Arbeit als Machine-Learning-Aufgabe die Regression bzw. Zeitreihenprognose gewählt, weshalb hier auf die Evaluation bei der Regression eingegangen wird. Hierfür stehen verschiedene Fehlermaße zur Verfügung, welche die Genauigkeit des Vorhersagemodells quantifizieren. Ausgewählte, gängige Fehlermaße und ihre Formeln zur Berechnung sind in Tabelle 2-1 abgebildet. Dabei steht $\hat{y}_i$ für den tatsächlichen Wert, y_i für den prognostizierten Wert und n für die Anzahl an prognostizierten Werten. Der MAE berechnet den durchschnittlichen absoluten Unterschied zwischen den tatsächlichen und den vorhergesagten Werten. Der MSE ist der Durchschnitt der quadrierten Unterschiede zwischen den tatsächlichen und vorhergesagten Werten. Der RMSE wird wiederum gebildet, indem aus dem MSE die Quadratwurzel gezogen wird. Maße wie der MSE und RMSE weisen eine hohe Ausreißersensitivität auf, da große Abweichungen überproportional gewichtet werden. Der MAE hingegen berücksichtigt nur die absolute Differenz und ist daher weniger empfindlich gegenüber Ausreißern. Der MAPE ist ein relatives, also prozentuales, Fehlermaß und eignet sich für zeitreihenübergreifende Vergleiche. Da hierbei die Abweichung der Vorhersage durch den Beobachtungswert dividiert wird, eignet sich diese Fehlerkennzahl jedoch nur für Zeitreihen, die ausschließlich positive Werte annehmen. Bei allen Fehlerkennzahlen gilt, je geringer die Werte sind, desto geringer sind die Abweichungen zwischen Prognose und Realität und desto besser prognostiziert das Modell die Zeitreihe (Mertens und Rässler 2012, 433–435).

Tabelle 2-1: Fehlerkennzahlen (Laaroussi et al. 2020, S. 3)

Name	Formel
Mean absolute error	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Mean square error	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Root mean square error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Mean absolute percentage error	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{\hat{y}_i}$

Der CRISP-DM Prozess endet mit dem *Deployment*. In dieser Phase wird strategisch der Einsatz und die Implementierung des Data-Mining-Modells geplant und umgesetzt (Chapmann et al. 2000, S. 32–34). Sollte das Ziel des Prozesses keine Implementierung eines Modells sein, kann sich die Phase auch auf das Erstellen Abschlussreports beschränken (Wirth und Hipp 2000b, S. 34–35). Das Deployment spielt in dieser Arbeit eine untergeordnete Rolle, da kein Modell in einem Prozess implementiert wird.

Der Aspekt der Datenqualität ist im Kontext der Erklärung der Datenvorverarbeitung adressiert und wird im Folgendem näher erläutert, da es sich um ein zentrales Konzept der Aufgabenstellung handelt.

2.1.3 Daten und Datenqualität

Um die Datenqualität eines gegebenen Datensatzes untersuchen zu können, wird eine geeignete Beschreibungsgrundlage benötigt. Im Feld der Datenanalyse greift die Untersuchung der Datenqualität auf die Verwendung von Dimensionen der Datenqualität zurück. Neben den Dimensionen der Datenqualität umfasst die Beschreibung der Daten das Identifizieren der Skalenniveaus der Attribute.

Die Skalenniveaus werden in drei Kategorien unterteilt. Zur ersten Kategorie zählen nominale Werte, welche sich durch eine rein beschreibende Funktion auszeichnen. Mit nominalen Werten sind keine Rechenoperationen möglich. Die ordinalen Werte ergänzen die nominalen um eine Rangfolge. Die letzte Kategorie bilden die kardinalen Werte, welche die Rangfolge um Abstände erweitern. Auf Basis dieser Abstände sind Rechenoperationen mit den Werten möglich (Bamberg et al. 2017, S. 6).

Untersuchungen der Qualität einer Datengrundlage basieren auf dem Ansatz, verschiedene Dimensionen der Datenqualität (DQ), welche jeweils einen Aspekt dieser abbilden, zu nutzen.

Eine der ersten umfassenden Zusammenstellungen der Dimensionen von DQ basiert auf einer Studie von Wang und Strong (1996). Im Rahmen der initialen Studie haben Wang und Strong (1996) 179 mögliche Eigenschaften der Datenqualität identifiziert und diese in einer Reihe von Meta-Dimensionen kategorisiert (Wang und Strong 1996, S. 16). Die so gefundenen Kategorien sind: „Intrinsic“, „Contextual“, „Representational“ und „Accessibility“. Diese Meta-Dimensionen umfassen jeweils eine Reihe von Eigenschaften, die die zusammengehörigen Dimensionen der Datenqualität darstellen. Unter der Kategorie „Intrinsic“ werden die Dimensionen, welche den Daten selbst zugesprochen werden, zusammengefasst. Hierzu gehören quantifizierbare Dimensionen, wie die Genauigkeit, aber auch Qualitative, wie die Glaubwürdigkeit und Objektivität der Daten. Unter der Kategorie „Contextual“ werden die kontextabhängigen Dimensionen zusammengefasst, wie die Aktualität oder die angemessene Menge der Daten. Eine Beurteilung dieser Dimensionen steht im Kontext zu einem betrachteten Anwendungsfall. Eigenschaften über die Darstellung der Daten werden in der Kategorie „Representational“ erfasst. Mit der letzten Kategorie „Accessibility“ wird die Zugänglichkeit oder auch die Zugriffsbeschränkung der Daten betrachtet (Batini und Scannapieca 2006, S. 38–42; Batini et al. 2009, S. 9; Lee et al. 2002, S. 16; Wang und Strong 1996, S. 16).

Aus der Menge an möglichen Dimensionen von Datenqualität werden in Tabelle 2-2 diejenigen beschrieben, die für die Untersuchung eines konkreten Datensatzes in einem KDD-Prozess relevant sind.

Tabelle 2-2: Dimensionen von Datenqualität (Auswahl nach Batini und Scannapieca (2006))

Name	Beschreibung	Quelle
Appropriate Amount of Data	Es liegt eine angemessene Menge an Daten vor, um den aktuellen Anwendungsfall zu bearbeiten.	(Batini und Scannapieca 2006, S. 38–42)
Timeliness	Die Daten sind für den Anwendungsfall zeitlich treffend.	(Batini und Scannapieca 2006, S. 38–42; Desai und Dinesha 2020, S. 1)
Value added	Beschreibt in welchem Grad die Daten einen Mehrwert für den zu bearbeitenden Anwendungsfall stiften.	(Batini und Scannapieca 2006, S. 38–42)

Name	Beschreibung	Quelle
Completeness	Beschreibt, ob die Daten sind für den Anwendungsfall hinreichend vollständig sind.	(Batini und Scannapieca 2006, S. 38–42; Blake und Mangiameli 2011, S. 4)
Accuracy	Untersucht ob und wie weit die Daten mit den Werten der realen Beobachtung übereinstimmen.	(Batini und Scannapieca 2006, S. 38–42; Blake und Mangiameli 2011, S. 4)

Konkret im Fall der Genauigkeit der Daten ist es weiter relevant zu betrachten, was für Fehler die Daten aufweisen können. Hier sind folgende Fehler hervorzuheben: Fehlende Werte, Ausreißer, inkonsistente Daten, Zeitinvarianz (Diop et al. 2017, S. 1). Im Allgemeinen werden zwei Ansätze zum Umgang mit solchen Fehlern vorgeschlagen, einmal posterior und prior, welche das Behandeln der Fehler im Datensatz und das Verhindern der Fehler im Voraus bedeuten (Diop et al. 2017, S. 2). Das Vorgehen der Fehlerbehebung setzt sich aus dem Detektieren und dem Beheben zusammen (Diop et al. 2017, S. 2). Der Schwerpunkt dieser Arbeit liegt auf dem Detektieren und Beheben von Ausreißern, daher wird im Weiteren der aktuelle Stand der Technik in diesem Themengebiet erarbeitet.

2.2 Durchführung der Literaturrecherche

Es wird eine Übersicht über gängige Datenvorverarbeitungsverfahren für die Korrektur von Ausreißern benötigt. Hierfür wird eine strukturierte Literaturrecherche durchgeführt, welche es ermöglicht, den Stand der Technik auf diesem Gebiet darzustellen. Das Vorgehen zur Durchführung einer strukturierten Recherche wird im Folgendem beschrieben. Im Anschluss wird die Recherche nach dem beschriebenen Vorgehen durchgeführt und die Ergebnisse dargestellt. Diese werden im Kontext der Aufgabenstellung diskutiert.

Die Durchführung der angestrebten Literaturrecherche gliedert sich in zwei Schritte. Zunächst wird das zu verwendende Vorgehensmodell dargestellt. Im Anschluss wird die konkrete Umsetzung der einzelnen Phasen des Vorgehens beschrieben.

2.2.1 Beschreibung des Vorgehensmodells

Die Durchführung der strukturierten Literaturrecherche basiert auf dem Vorgehensmodell von vom Brocke et al. (2009) aus dem Jahr 2009. Dieses wird in diesem Abschnitt zunächst zusammengefasst, um im Anschluss die konkrete Instanz für die durchzuführende Recherche zu formulieren. Eine strukturierte Literaturrecherche beschreibt nach vom Brocke et al. (2009) die Art von Recherche, die einer klaren Methodik folgt, weshalb der Begriff im Rahmen der hier angestrebten Ausarbeitung verwendet wird, auch wenn es sich nicht zwangsläufig um eine vollumfängliches Literaturreview handelt (vom Brocke et al. 2015, S. 208; vom Brocke et al. 2009, S. 9).

Das von vom Brocke et al. (2009) beschriebene Vorgehensmodell gliedert sich in fünf Phasen, welche sequenziell durchlaufen werden. In der ersten Phase wird der Anwendungsbereich des Reviews festgelegt. Die Beschreibung dieses Anwendungsbereich oder auch Geltungsbereich einer Recherche geht auf eine Veröffentlichung von Cooper (1988) aus dem Jahr 1988 zurück. Dort beschreibt Cooper eine Taxonomie, mit der der Geltungsbereich eines Reviews festgelegt werden kann. Diese Taxonomie basiert auf sechs Dimensionen, die in Tabelle 2-3 dargestellt sind. Die Taxonomie wird von vom Brocke et al. (2015) weiter um die in Tabelle 2-5 dargestellten Aspekte ergänzt. Auf Basis dieser Taxonomien von Cooper (1988) und vom Brocke et al. (2015) kann der Geltungsbereich und der Umfang einer angestrebten Recherche festgelegt werden (vom Brocke et al. 2009, S. 8–9; vom Brocke et al. 2015, S. 214; Cooper 1988, S. 109).

In der zweiten Phase findet im Modell nach vom Brocke et al. (2009) die Konzeptualisierung des Themas statt, hierbei sollen die Schlüsselbegriffe des Themenfeldes erarbeitet werden. Als Grundlage hierfür können nach vom Brocke et al. (2009) Lehrbücher oder Enzyklopädien genutzt werden. Das Ergebnis dieser Phase sollten die im weiteren benötigten Schlüsselbegriffe sowie deren Synonyme sein (vom Brocke et al. 2009, S. 10).

In der dritten Phase beschreibt vom Brocke et al. (2009) wie die Literatur erfasst wird. Hierbei wird zunächst festgelegt welche Datenbanken und welche Schlüsselwörter und Schlüsselwortkombinationen genutzt werden. Darüber hinaus wird festgelegt, ob die Nutzung von Vorwärts und oder Rückwärtssuchen durchzuführen ist. Außerdem gilt es zu bestimmen welche Art von Veröffentlichungen in Betracht gezogen werden (vom Brocke et al. 2009, S. 10–11).

In der vierten Phase verorten vom Brocke et al. (2009) eine Analyse und Synthese der Ergebnisse aus der dritten Phase. Die vorgeschlagene Methode hierfür ist die Konzeptmatrix nach Webster und Watson (2002). Eine Konzeptmatrix ist eine konzeptzentrierte Darstellung von Aspekten aus der Literatur, die zu gemeinsamen Kategorien oder auch Dimensionen zusammengefasst werden. Neben der konzeptzentrierten Sicht existiert weiter die autorenspezifische Sicht, welche sich nach Webster und Watson (2002) jedoch nicht dazu eignet, ein Themengebiet zu erfassen. Eine weitere Möglichkeit der Darstellung von Rechercheergebnissen ist die Entwicklung einer eignen Taxonomie, wie Nickerson et al. (2009) in Form eines entsprechenden Vorgehensmodells beschrieben haben (Webster und Watson 2002, S. 17; vom Brocke et al. 2009, S. 8–9, 2009, S. 11; Nickerson et al. 2009, S. 6).

Abgeleitet aus den Ergebnissen ergeben sich für Vom Brocke in der fünften Phase Fragestellungen für weitere Forschungsansätze (vom Brocke et al. 2009, S. 11).

Tabelle 2-3: Taxonomie eines Literatur Reviews nach (Cooper 1988, S. 109)

Dimensions	Categories			
Focus	Research Outcomes	Research Methods	Theories	Practice or Applications
Goal	Integration	Criticism	Identification of Central Issues	
Perspective	Neutral Representation		Espousal of Position	
Coverage	Exhaustive	Exhaustive with Selective Citation	Representation	Central or Pivotal
Organization	Historical	Conceptual	Methodological	
Audience	Specialized Scholars	General Scholars	Practitioners or Policy Makers	General Public

Tabelle 2-4: Definition des Scopes nach (vom Brocke et al. 2015, S. 214)

Dimensions	Categories		
Process	Sequential	Iterative	
Sources	Citation indexing services	Bibliographic databases	Publications
Coverage	Comprehensive	Representative	Seminal works
Techniques	Keyword Search	Backward search	Forward search

Im nächsten Schritt wird das beschriebene formale Vorgehensmodell umgesetzt, beginnend mit der ersten Phase, dem Festlegen des Umfangs und des Geltungsbereiches der Recherche.

2.2.2 Festlegen des Scopes der Recherche

Basierend auf den Taxonomien von Cooper (1988) und vom Brocke et al. (2015), dargestellt in Tabelle 2-3 und Tabelle 2-4, gilt es, aus den dort beschriebenen Dimensionen die für diese Recherche relevanten Ausprägungen zu bestimmen, um auf diese Weise den Umfang und Geltungsbereich der Recherche festzulegen. Die gewählten Ausprägungen sind in Tabelle 2-5 zusammengestellt. Aus der in der Einleitung dargestellten Aufgabenstellung dieser Arbeit ist der Umfang der Recherche abzuleiten. Die hier durchzuführende Recherche verfolgt den Zweck, die im heutigen Stand der Technik möglichen und nützlichen Methoden für die Detektion und das Beheben von Ausreißern zu identifizieren. Aus dieser Aufgabenstellung lassen sich im nächsten Schritt die Ausprägungen der gewählten Taxonomie über den Umfang der Recherche ableiten. Die Abdeckung bzw. der Umfang der Recherche soll repräsentativ sein. Der Fokus der Recherche liegt auf den in den Veröffentlichungen beschriebenen Ergebnissen sowie den aufgezeigten Anwendungsmöglichkeiten im Einsatz von Ausreißererkennungsverfahren. Die Recherche wird als iterativer Prozess durchgeführt, um so gegebenenfalls bestimmte Aspekte zu vertiefen oder bei zu geringer Trefferanzahl einer Keyword-Suche zusätzlich eine Rückwärts- oder Vorwärtssuche zu nutzen. Verwendet werden Quellen, die als Veröffentlichung einem Peer-Review unterzogen worden sind, wie Konferenzbeiträge und Veröffentlichungen in Journalen. Das Ziel der Recherche liegt in der Identifikation von zentralen Aspekten und Methoden und soll sich am Ende an ein Publikum aus Wissenschaftler verschiedener Fachbereiche richten.

Tabelle 2-5: Festlegen des Scopes der Recherche basierend auf den Taxonomien von Cooper 1988 und Vom Brocke et al. 2015

Dimension	Ausprägung
Abdeckung	Repräsentative
Fokus	Ergebnisse und Anwendungen
Prozess	Iterative
Quellen	Geprüfte Veröffentlichungen
Techniken	Primär Schlüsselwort Suche, abweichend Vor- und Rückwärtssuche
Ziel	Identifikation von zentralen Aspekten
Zielgruppe	Allgemeine und spezialisierte Wissenschaftler

Im nächsten Schritt sind nun die fachlichen Konzepte zu beschreiben, welche die Grundlage der Recherche bilden. Diese können der Empfehlung von Vom Brocke folgend aus Sach- und Lehrbüchern herausgearbeitet werden. Sobald alle benötigten Konzepte eingeführt sind, kann mit der Phase des Datensammelns begonnen werden. Als fachliche Grundlage wird hier weiter das bereits eingeführte Vorgehensmodell CRISP-DM genutzt. Das Vorgehensmodell CRISP-DM bietet nun eine Struktur, anhand derer die Suche eingegrenzt werden kann. Die Fragestellung der Identifikation und Behebung von Ausreißern ist im Data Understanding und der Datenvorverarbeitung angesiedelt.

Es folgt die dritte Phase, das Sammeln der Daten. Hierfür wird die Fachdatenbanken IEEE mit einem ingenieurswissenschaftlichen Schwerpunkt genutzt, sowie Scopus und Scince Direct mit einer vollständigen Abdeckung an Themengebieten.

Die Ergebnisse der Recherche werden gefiltert. In diesem Schritt werden alle Veröffentlichungen verworfen, die die gesuchten Schlüsselworte enthalten, diese aber nicht im gefragten richtigen Kontext betrachten. Es werden weiter alle Veröffentlichungen gefiltert, die nicht zugänglich sind.

Die relevante Literatur ist in Tabelle 7-1 in Form einer Konzeptmatrix dargestellt. Neben den verschiedenen Ansätzen zur Ausreißererkennung sind Konzepte der Ausreißerbeurteilung und Ansätze zur Behebung aufgeführt. Die Erkennungsverfahren werden im nächsten Abschnitt im Detail untersucht.

2.3 Analyse der Rechercheergebnisse

Während der Literaturrecherche gewonnene Forschungsergebnisse werden in Abbildung 2-2 als morphologischer Kasten visualisiert. Die Anordnung im morphologischen Kasten orientiert sich an den Phasen des CRISP-DM-Modells. In der Datenvorverarbeitungsphase liegt der Fokus besonders auf der Verbesserung der Datenqualität. Dabei werden spezifische Verfahren zur Erkennung von Ausreißern identifiziert und mögliche Ansätze zu ihrer Behebung untersucht. Da diese Arbeit sich auf die Behandlung von Ausreißern konzentriert, wird im Verlauf dieses Kapitels eine detaillierte Untersuchung der Ausreißer durchgeführt. Zuerst werden die Grundlagen zu Ausreißern erklärt. Anschließend werden verschiedene Methoden zur Erkennung von Ausreißern aufgelistet und mithilfe von Grafiken näher erläutert. Zum Abschluss werden mögliche Verfahren zur Behandlung von Ausreißern präsentiert.

Business Understanding		Vorgehensmodelle	CRISP-DM	SEMMA	KDD	Weitere	
		DM-Aufgabe	Klassifikation	Clustering	Regression	Wirkzusammenhänge	
Data Understanding		Datentypen	Ganzzahl	Fließkommazahl	Datum	Zeichenkette	Boolean
		Skalenniveaus	Nominal	Ordinal	Kardinal		
Data Preparation	Datenqualität	Dimensionen von DQ	Genauigkeit	Vollständigkeit	Aktualität	Konsistenz	Und Weitere
		Fehlerarten	Ausreißer	Fehlenden Werte	Rauschen	Inkonsistenzen	
		Umgang mit Fehlern	Löschen	Beheben	Ignorieren		
		Ausreißer Identifikation	Distanz	Statistisch	Machine Learning	Dichte	Cluster
		Ausreißer beheben	Imputation	Glättung	Löschen		
		Daten Reduktion	Zeilen	Spalten	Wertebereich		
Modeling		ML-Modelle	Überwacht	Unüberwacht	Verstärkend		
Evaluation		Bewertungsmethoden	Performancemaße (Präzision, Sensitivität, Spezifität, F-Maß, Lift)	Konfusionsmatrix	Kreuzvalidierung	ROC-Kurve	

Abbildung 2-2: Ergebnisse der Literaturrecherche (eigene Darstellung)

In der Data-Mining- und Statistik-Literatur werden Ausreißer auch als Anomalien, Diskordanten oder Abweichungen bezeichnet (Aggarwal 2017, S. 1; Chandola et al. 2009, S. 2). Laut Hawkins (Hawkins 1980, S. 1) wird eine Beobachtung als Ausreißer betrachtet, wenn sie signifikant von den anderen Beobachtungen abweicht und den Verdacht aufkommen lässt, dass sie durch einen anderen Mechanismus entstanden ist. Ausreißer weichen daher vom normalen oder bekannten Verhalten der Daten ab und weisen Werte auf, die weit von den erwarteten oder durchschnittlichen Werten entfernt sind oder keine Verbindung zu anderen Objekten in Bezug auf ihre Merkmale aufweisen (Ranga Suri et al. 2019, S. 3; Chandola et al. 2009, S. 2–3; Ferdowsi et al. 2013, S. 1).

Die Frage, was eine ausreichende Abweichung darstellt, um ein Objekt als Ausreißer zu identifizieren, ist oft subjektiv (Aggarwal 2017, S. 2). Daher gestaltet sich die Definition und Abgrenzung von Ausreißern meistens als schwierig (Chandola et al. 2009, S. 3).

In den meisten Anwendungen werden Daten durch einen oder mehrere Generierungsprozesse erstellt, die entweder die Aktivitäten im System widerspiegeln oder Beobachtungen über Entitäten erfassen. Wenn der Generierungsprozess ungewöhnlich verläuft, führt dies zur Entstehung von Ausreißern (Aggarwal 2017, S. 1, 2015, S. 17, 2015, S. 17; Ferdowsi et al. 2013, S. 1). Daher enthält ein Ausreißer oft nützliche Informationen über abnormale Eigenschaften der Systeme und Entitäten, die den Prozess der Datengenerierung beeinflussen und zum Beispiel die Ergebnisse der Datenanalyse verzerren können (Carmona et al. 2020, S. 589). Die Erkennung solcher ungewöhnlichen Charakteristika ermöglicht wertvolle, anwendungsbezogene Erkenntnisse (Aggarwal 2017, S. 1).

Im Unterschied zum Rauschen handelt es sich bei Ausreißern in der Regel um bedeutsame Abweichungen, die von besonderem Interesse sind (Aggarwal 2017, S. 2–3). Rauschen bildet in unüberwachten Szenarien die semantische Grenze zwischen normalen Daten und Ausreißern. Es wird oft als eine mildere Form von Ausreißern betrachtet (Ranga Suri et al. 2019, S. 8). Die Identifikation und Entfernung von Rauschpunkten stellen dennoch eine bedeutende Aufgabe im Data Mining dar. Daher sind sowohl das Rauschen als auch die Probleme bei der Erkennung von Ausreißern gleichermaßen wichtig für eine effektive Datenanalyse (Ranga Suri et al. 2019, S. 8).

Dieser Bericht fokussiert sich auf die Erkennung und Behandlung von Ausreißern. Beide Schritte sind entscheidend und müssen der Datenanalyse vorausgehen. Das Ziel besteht darin, abnormale Daten von den normalen Daten im Datensatz zu separieren. Ausreißer treten in der Regel isoliert oder kontinuierlich in den Daten auf (Wan et al. 2019, S. 173827).

Die Ergebnisse eines Ausreißererkennungsalgorithmus können auf zwei Arten präsentiert werden. Einerseits generieren die meisten Ausreißererkennungsalgorithmen sogenannte "Scores", die die Abweichungen jedes Datenpunkts quantifizieren. Diese Scores ermöglichen es, die Datenpunkte nach ihrer Neigung zu Ausreißern zu sortieren (Aggarwal 2017, S. 2). Andererseits können auch binäre Labels als alternative Form der Ausgabe verwendet werden, um anzuzeigen, ob ein Datenpunkt als Ausreißer identifiziert wurde. Scores liefern detaillierte Informationen über Abweichungen, während binäre Labels eine einfache Grundlage für praktische Entscheidungsfindung bieten (Aggarwal 2017, S. 2).

Zusammenfassend gibt es keine allgemeingültige Definition für Ausreißer, weshalb es schwierig ist, in einer Menge von Datenobjekten eine bestimmte Anzahl von Objekten zu finden, die sich von den übrigen Daten erheblich unterscheiden und sowohl außergewöhnlich als auch inkonsistent sind (Ranga Suri et al. 2019, S. 15). Außerdem gibt es viele Erkennungsmethoden, die für verschiedene Bereiche geeignet sind und vom Datensatz abhängig sind (Wan et al. 2019, S. 173827; Aggarwal 2017, S. 7).

In der Literatur gibt es folgende Gruppen von Verfahren: Statistische Verfahren, distanzbasierte Verfahren, Machine-Learning-Verfahren, dichtebasierte Verfahren, verteilungsbasierte Verfahren und clusterbasierte Verfahren (Angelov et al. 2015, S. 524–526; Dost et al. 2017, S. 220; Ferdowsi et al. 2013, S. 1; Han et al. 2008, S. 97; Imtiaz Ahmed et al. 2018, 2018; Jayaramulu und Venkateswarlu 2022, S. 1143–1144; Wan et al. 2019, S. 173828). Der Fokus der weiteren Ausarbeitung liegt bei den meistverwendeten Verfahren. Diese sind die statistischen, distanzbasierten, clusterbasierten und dichtebasierten Verfahren. Sie lassen sich in die Hauptkategorien statistische und näherungs- bzw. distanzbasierte Ansätze gruppieren. Diese Beziehungen werden auch durch die Abbildung 2-3 zusammengefasst (Ranga Suri et al. 2019, S. 15).

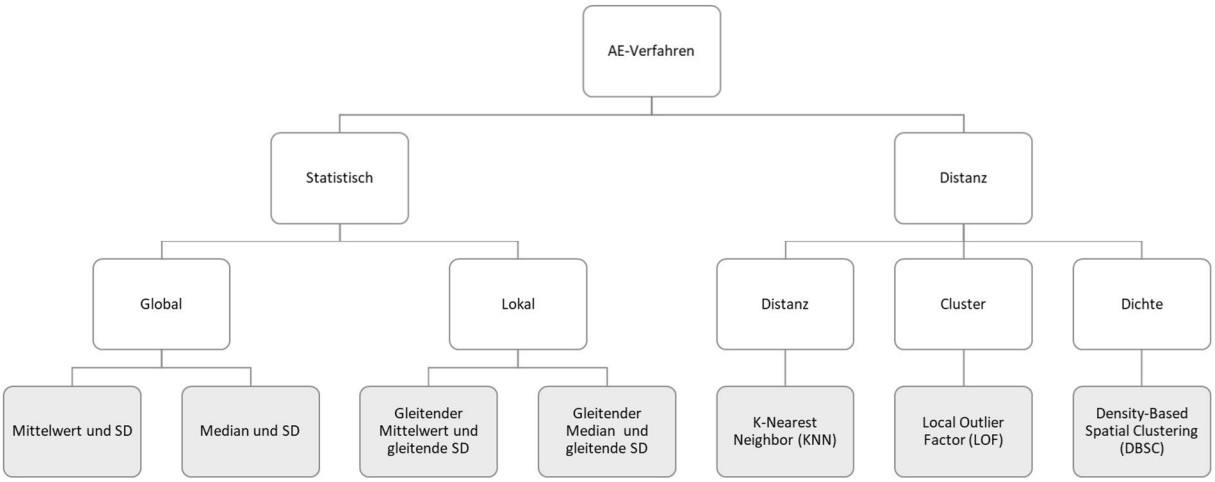

Abbildung 2-3: Übersicht der genutzten Ausreißererkennungsverfahren (eigene Darstellung)

2.3.1 Statistische Ausreißererkennungsverfahren

Auf der linken Seite der Abbildung 2-3 sind die statistische Verfahren aufgeführt, welche sich weiter in globale und lokale Verfahren unterteilen lassen (Pearson 2002, S. 59; Muthukrishnan et al. 2004, S. 41).

Ein statistischer Ansatz geht von einem Verteilungs- oder Wahrscheinlichkeitsmodell für die gegebenen Daten aus und identifiziert Ausreißer in Bezug auf dieses Modell mithilfe von Diskordanz-Tests (Ranga Suri et al. 2019, S. 15). Demnach sind Ausreißer Beobachtungen, die statistisch gesehen nicht mit den übrigen Daten übereinstimmen (Ranga Suri et al. 2019, S. 30). Normale Datenpunkte folgen der Verteilungsmethode, welches auch der Abbildung 2-4 zu entnehmen ist (Mandhare und Idate, Prof. S. R. Idate 2017, S. 933). In der Abbildung 2-4 stellt der Bereich, in dem sich die grünen Datenpunkte befinden, den Toleranzbereich dar. Sobald sich ein Datenpunkt nicht in diesem Bereich befindet, gilt dieser als Ausreißer (hier die roten Datenpunkte).

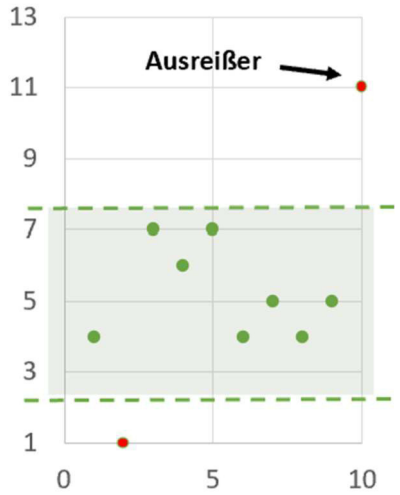

Abbildung 2-4: Statistisch basierte AE (eigene Darstellung)

Es hat sich jedoch gezeigt, dass die modellbasierte Erkennung von Ausreißern verschiedene praktische Grenzen hat. Um den mit den statistischen Methoden verbundenen Einschränkungen zu begegnen, werden näherungs- bzw. distanzbasierte Ansätze eingeführt (Ranga Suri et

al. 2019, S. 15). Diese Ansätze sind in der Abbildung 2-3 auf der rechten Seite dargestellt. Diese Methoden beruhen auf einem wohldefinierten Begriff des Abstands, um den Abstand zwischen einem Paar von Datenobjekten zu messen (Ranga Suri et al. 2019, S. 32–33). Demnach klassifizieren näherungsbasierte Methoden einen Datenpunkt als Ausreißer, wenn dieser in einem dünn besiedelten Bereich liegt oder wenige Nachbarn in seiner Umgebung hat. Die Nähe zwischen den Datenpunkten wird anhand ihrer geringfügigen Unterschiede definiert. Dennoch sind sie ähnlich genug, um in einer Gruppierung zusammengefasst zu werden. Die gängigsten Ansätze zur Bestimmung dieser Nähe für die Ausreißeranalyse umfassen distanzbasierte Verfahren, clusterbasierte Verfahren sowie dichtebasierte Verfahren (Carmona et al. 2020, S. 590; Bhattacharya et al. 2015, S. 24).

2.3.2 Distanzbasierte Ausreißererkennungsverfahren

Die distanzbasierten Verfahren werden als gängigste Technik zur Ausreißererkennung gesehen. Bei diesen wird der Abstand eines Datenpunktes von seinen Nachbarn berechnet. Die Objekte, die sich in einem größeren Abstand zu ihrer Gruppe befinden, sind Ausreißer (Mandhare und Idate, Prof. S. R. Idate 2017, S. 933). Dementsprechend werden distanzbasierte Methoden basierend auf den Konzepten der lokalen Nachbarschaft oder k-nächsten Nachbarn (kNN) der Datenpunkte definiert (Carmona et al. 2020, S. 590). Der euklidische Abstand zum k-nächsten Nachbarn ist dabei die am meisten bevorzugte Methode (Mandhare und Idate, Prof. S. R. Idate 2017, S. 933). Dies ist auch in der Abbildung 2-5 zu erkennen. Der rote Datenpunkt stellt den Ausreißer dar, da der Abstand zu dem k-nächsten Nachbarn größer ist. Die grünen Datenpunkte stehen für die „normalen“ Daten.

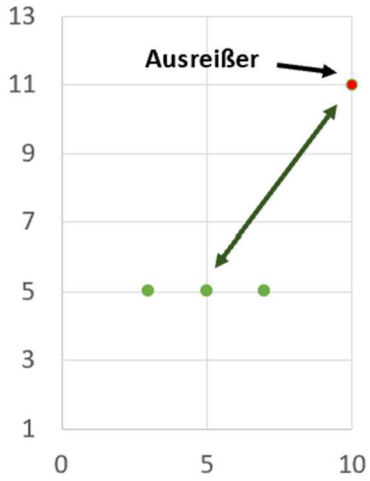

Abbildung 2-5: Distanzbasierte AE (eigene Darstellung)

2.3.3 Clusterbasierte Ausreißererkennungsverfahren

Eine clusterbasierte Methode gruppiert Datenobjekte basierend auf den Informationen, die aus den Objekten und ihren Beziehungen gewonnen werden. Ziel ist es, ähnliche Datenpunkte zusammenzufassen (Mandhare und Idate, Prof. S. R. Idate 2017, S. 933; Ranga Suri et al. 2019, S. 36–37).

Die Auswahl der Clustering-Methoden richtet sich nach der Anwendung, da diese Methoden unterschiedliche Komplexitäten aufweisen, um sich an Cluster mit variierenden Anzahlen, Größen und Formen anzupassen (Gupta et al. 2014, S. 10). Dabei liegt das Hauptaugenmerk auf der Bestimmung der Clusteranzahl. Ein clusterbasierter Ansatz und ein distanzbasierter Ansatz führen zum gleichen Ergebnis, wenn nur ein Cluster existiert (Mandhare und Idate, Prof. S. R. Idate 2017, S. 933; Ranga Suri et al. 2019, 36–37).

Ein wichtiger Wert in clusterbasierten Ausreißererfassungsverfahren ist der "lokale Ausreißerfaktor" (LOF). Dieser berücksichtigt die Dichte eines Datenpunkts im Vergleich zu seinen nächsten Nachbarn. Bei dieser Methode wird überprüft, ob ein Objekt von seinen Nachbarn isoliert ist, basierend auf seinem lokalen Ausreißerfaktor. Objekte mit dem höchsten LOF gelten als Ausreißer, während solche mit niedrigem LOF als normale Datenpunkte betrachtet werden (Mandhare und Idate, Prof. S. R. Idate 2017, S. 934; Breunig et al. 2000, S. 94–97).

In Abbildung 2-6 sind grüne Datenpunkte zu erkennen, die zwei Cluster bilden. Im Gegensatz dazu hat der rote Datenpunkt niedrige LOF-Werte und gilt als Ausreißer.

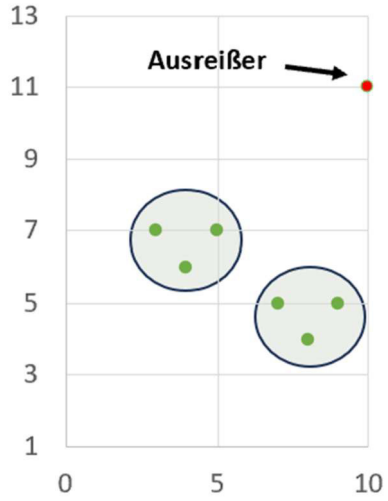

Abbildung 2-6: Clusterbasierte AE (eigene Darstellung)

2.3.4 Dichtebasierte Ausreißererkennungsverfahren

Die dichtebasierten Methoden zur Ausreißererkennung vergleichen die Dichte eines spezifischen Datenpunkts mit jener seiner benachbarten Datenpunkte. Dieser Dichteunterschied wird in Form eines Ausreißer-Scores erfasst. Während normale Datenpunkte und ihre Nachbarn gleiche Dichten haben, weisen Ausreißer abweichende Dichten auf (Mandhare und Idate, Prof. S. R. Idate 2017, S. 933).

Der DBSCAN-Clustering-Algorithmus wird verwendet, um räumliche Cluster zu identifizieren und potenzielle räumliche Ausreißer zu finden. Wenn Cluster unterschiedliche Dichten haben, wird jedem Cluster ein Dichtefaktor zugewiesen. Dieser Faktor wird mit dem Durchschnittswert eines Clusters verglichen, sobald ein neuer Datenpunkt eintrifft. Nach dem Clustering werden potenzielle räumliche Ausreißer ermittelt. Dies erfolgt durch die Überprüfung der räumlichen Nachbarn, um festzustellen, ob diese Objekte tatsächlich räumliche Ausreißer sind (Ester et al., S. 226–228).

Die folgende Abbildung 2-7 zeigt die grünen Datenpunkten mit einer gleichen Dichte, während diese von der berechneten Dichte des roten Datenpunktes abweicht, der somit als Ausreißer bezeichnet wird.

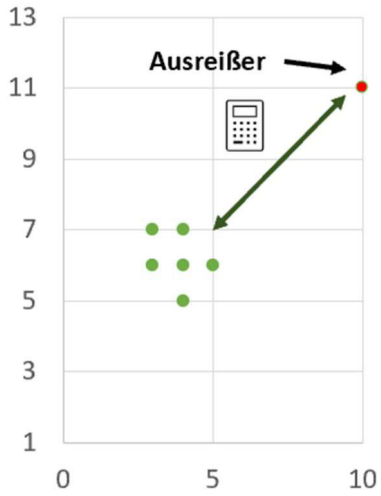

Abbildung 2-7: Dichtebasierte AE (eigene Darstellung)

2.3.5 Behebung von Ausreißern

Nachdem Ausreißer entdeckt wurden, steht die Entscheidung an, wie mit ihnen umgegangen werden soll. Es gibt verschiedene Techniken zur Ausreißerbehandlung, die verwendet werden können, um Ausreißer oder abnormale Datenpunkte in einem Datensatz zu behandeln oder zu korrigieren. Dazu gehören Glättungs- und Imputationsverfahren (Chen et al. 2021, S. 341). Das Löschen von Ausreißern ist ebenfalls eine Möglichkeit zur Ausreißerbehebung (Chen et al. 2021, S. 341). Dies kann jedoch dazu führen, dass wertvolle Informationen verloren gehen, besonders wenn die Ausreißer tatsächlich wichtige Informationen repräsentieren (Osborne und Overbay 2004, S. 3).

Glättungsverfahren sind statistische Techniken, die dazu dienen, Muster und Trends in den Daten zu identifizieren, während Ausreißer reduziert werden. Ausreißer können mithilfe von Methoden wie der Mittelwert- oder Medianberechnung korrigiert werden (Chen et al. 2021, S. 341). Wenn Ausreißerwerte fehlende oder ungültige Werte darstellen, können sie mittels Imputationsverfahren mit angemessenen Werten ergänzt oder ersetzt werden (García et al. 2015, S. 13; Chheda et al. 2021, S. 28; Aggarwal 2015, S. 35, 2015, S. 28; Moreno-Sanchez 2020, S. 3787).

Insgesamt zielt die Forschung darauf ab, Ausreißer in den Daten zu identifizieren, da diese Abweichungen von normalen Mustern wichtige Informationen liefern können. Unterschiedliche Methoden wie statistische, distanzbasierte, clusterbasierte und dichtebasierte Ansätze werden verwendet, um solche Ausreißer zu erkennen. Die Behandlung von Ausreißern erfordert kontextabhängige Entscheidungen, bei denen Glättungs- und Imputationsverfahren sowie das Löschen von Ausreißern angewendet werden können. Die Wahl der richtigen Methode hängt von verschiedenen Faktoren ab, wie z.B. dem Datenkontext. Generell ist die Identifizierung und Behandlung von Ausreißern innerhalb der Datenanalyse eine essenzielle Aufgabe, die ein genaues Verständnis und einen angemessenen Ansatz erfordert.

3 Untersuchung der Verfahren zur Behandlung von Ausreißern folgend dem CRISP-DM Vorgehensmodell

Im Rahmen dieses Kapitels soll der CRISP-DM Prozess für eine konkrete Data-Mining-Fragestellung bearbeitet werden. Das Ziel hierbei ist es, Ergebnisse zu liefern, die die Untersuchung des Einflusses der Ausreißer auf die Data-Mining-Fragestellung ermöglichen. Das Kapitel ist in drei Teilen strukturiert. Im ersten Teil 3.1 wird das Data und Business Understanding bearbeitet und eine Data-Mining-Fragestellung entwickelt. Der zweite Teil 3.2 beschreibt die Umsetzung der Datenvorverarbeitung sowie die Modellentwicklung. Im letzten Teil 3.3 wird die Implementierung der Ausreißererkennungsverfahren dargestellt.

3.1 Ableiten einer Data Mining Fragestellung aus dem Data Understanding

Die Herausforderung der hier untersuchten Aufgabenstellung besteht darin, anhand der Produktionsdaten realer Prozesse Verfahren der AE zu nutzen und Rückschlüsse über den Einfluss der Ausreißer auf die Data-Mining-Ergebnisse zu erzielen, ohne dass eine initiale Fragestellung an die Daten besteht. Daraus folgt, dass im Rahmen des Data Understandings der Datensatz auf eine geeignete Fragestellung untersucht wird. Diese geeignete Fragstellung wird dann im Rahmen des CRISP-DM-Prozesses mithilfe von Data-Mining-Verfahren bearbeitet. Dies bildet so einen Rahmen für die Untersuchung der verschiedenen AE-Verfahren im Kontext dieser konkreten Fragstellung.

3.1.1 Beschreibung der Daten

Ein erster Versuch, sich einem Datensatz im Rahmen des Data Understandings zu nähern, kann darin bestehen, diesen mit geeigneten Datenanalyse-Methoden beschreiben zu lassen. Für den hier vorliegenden Datensatz ergeben sich eine Tupel-Anzahl von 10 Millionen und neun verschiedene Attribute. Es fällt auf, dass die Attribute keinen einheitlichen Datentypen aufweisen. Um ein besseres Verständnis über die Daten zu gewinnen, bietet es sich an, ein Data Dictionary zu erstellen. Ein solches Data Dictionary ist für einen Ausschnitt der Daten in Tabelle 3-1 abgebildet. Das Data Dictionary bietet einen Überblick über die Datentypen und Skalenniveaus der einzelnen Attribute. Zusätzlich kann eine statistische Auswertung der Daten ergänzt werden. Da es sich im vorliegenden Fall vermehrt um nominale Daten handelt, beläuft sich die Auswertung auf die Anzahl an Kategorien und deren häufigster Werte. Neben der formalen Betrachtung werden die Daten inhaltlich untersucht, dafür werden im Weiteren die einzelnen Attribute oder Attributkombinationen erläutert.

Das Attribut WorkpieceGuid (WpG) ist ein identifizierendes Merkmal der einzelnen im System bearbeiteten Werkstücke. Im gewählten Ausschnitt des Data Dictionarys liegen 11158 verschiedenen Kategorien des Attributes vor, was bedeutet, dass genau so viele verschiedene Werkstücke im Datensatz beschrieben werden. Die Attribute WorkSequenz (WS), RoutingSEquence (RoS) und ResultSequence (ReS) beschreiben die Kombination aus Arbeitsschritt, Arbeitssequenz und Arbeitsplatz, an dem der vorliegende Datenpunkt aufgenommen wurde. Dies erklärt die ordinale Skalierung der Attribute. Die Arbeitsschritte sind fortlaufend nummeriert und werden sequenziell durchgeführt, daher befinden sie sich in einer Rang- oder Reihenfolge, jedoch sind mathematische Operationen an diesen Daten nicht zulässig. Das nächste Attribut ist die ParamterDescriptionId (PDID). Die PDID beschreibt die Art des Aufgenommenen Parameter am vorliegenden Arbeitsschritt. Der aufgezeichnete Wert ist im Attribut Value hinterlegt. Hier ist zu beachten, dass für die unterschiedlichen PDIDs verschiedene Wertebereiche und Skalenniveaus in Frage kommen, daher kann für das Attribut Value kein einheitlicher Skalen- oder Wertebereich festgelegt werden. Im Attribut Result wird das Ergebnis der vorgenommenen Messung bewertet. Für die Bewertung stehen fünf verschiedene Kategorien zur Verfügung. Im Attribut TimeStamp (TS) wird der Zeitpunkt der Messung aufgezeich-

net. Die letzten beiden Attribute sind die ProductID (PID) und der Wochentag. Die PID schlüsselt die sechs verschiedenen Produkte auf. Mit dem Attribut Wochentag wird der Wochentag angegeben, an dem der Wert aufgezeichnet wurde.

Tabelle 3-1: Data Dictionary für einen Datenausschnitt (eigene Darstellung)

Attribut	Datentyp	Skalenniveau	Häufigster Wert	Anzahl an Kategorien
WorkpieceGuid	String	Nominal	51AC70B2-CB1B-47D8-AAFA-00E6A95C6162	11158
WorkSequence	Integer	Ordinal	2	50
RoutingSequence	Integer	Ordinal	20	9
ResultSequence	Integer	Ordinal	1	56
ParameterDescriptionId	String	Nominal	PDES00000164, PDES00000179	76
Value	Object			
Result	String	Nominal	Pass	5
TimeStamp	Date Time	Kardinal		
ProductId	String	Nominal	PROD00000006	6
Weekday	Integer	Ordinal	2	7

3.1.2 Erarbeiten der Fragestellung

In den weiteren Untersuchungen der Daten gilt es, eine für die Ausreißerdetektion geeignete Data-Mining-Fragestellung zu entwickeln. Die Ausprägungen „Fail“ und „Pass“ des Attributes Result lassen darauf schließen, dass im Prozess Störungen auftreten, die das Ergebnis der Produktionsschritte beeinflussen. Die Ausprägungen des Attributes Value zeigen, dass die Bewertung des Ergebnisses auf den gemessenen Wert zurückzuführen ist. Der Einsatz eines Entscheidungsbaumes zur Klassifikation des Attributes Results bestätigt diese Annahme, da ein regelbasierter Zusammenhang zwischen dem Attribtut Value in Kombination mit den Attributen der Arbeitsschritte sowie dem Result hergestellt werden kann. Hieraus lässt sich schließen, dass die Bewertung auf einem klaren Regelwerk innerhalb des Unternehmens basiert. Eine Prognose zur Klassifikation des Attributes Result liefert folglich im Rahmen einer Data-Mining-Fragestellung kein neues Wissen. Diese Erkenntnis wirft die Frage auf, ob die Prognose des Attributes Value der konkreten Werte für zukünftige Zeitpunkte möglich ist. Eine Prognose zukünftiger Werte des Attributes Value würde nützliche Informationen darstellen, da diese, wenn sie korrekt prognostiziert sind, Informationen über zukünftige Ausfälle liefern. Um dieser Fragestellung nachzugehen, muss ein Unterraum in den Daten identifiziert werden, der die Prognose zukünftiger Werte des Attributes Value ermöglicht und gleichzeitig den Einsatz der verschiedenen Ausreißererkennungsverfahren zulässt. Im nächsten Schritt werden die Daten visualisiert, um einen solchen Unterraum zu finden. Es zeigt sich, dass die Werte des Attributes Value eine einheitliche Form annehmen, sobald man diese entsprechend der PDID filtert. Dies unterstreicht, dass die PDID den Parametertypen der Messung klassifiziert.

In Abbildung 3-1 ist ein Histogramm über die Häufigkeit der einzelnen PDIDs dargestellt: 90% der verschiedenen PDIDs treten knapp eine Million Mal auf und der verbleibende Teil von 10% mehr als zwei Millionen Mal. Diese annähernde Gleichverteilung der PDIDs legt für das weitere Vorgehen den Schluss nahe, dass für jede einzelne PDIDs ausreichend Datenpunkte für eine Betrachtung gegeben sind.

Im Fall der verschiedenen Produkte (gegeben durch die Produkt-IDs), auch in Abbildung 3-1 dargestellt, lässt sich erkennen, dass für die Produkte „011“ und „014“ bedeutend weniger Datenpunkte vorliegen im Gegensatz zu den anderen gezeigten Produkten.

Für die Prognose von nützlichen Werten ist die Betrachtung der Verteilung des Attributes Result relevant. Im betrachteten Fall einer Prognose von Werten, die zu verschiedenen Ausprägungen des Attributes Result führen, ist es unerlässlich, dass der gesuchte Unterraum des Datensatzes auch verschiedene Ausprägungen dieses Attributes enthält. Wird die Verteilung der verschiedenen Ausprägungen in Abbildung 3-2 betrachtet, ist zu erkennen, dass eine starke Imbalance zwischen den verschiedenen Ausprägungen vorliegt. Die Ausprägung „PASS“ ist knapp um den Faktor tausend häufiger vertreten als die Ausprägung „FAIL“. Wird die Aussagekraft der einzelnen Ausprägungen betrachtet, ist davon auszugehen, dass besonders „PASS“ und „FAIL“ die gefragten Zustände im Produktionssystem darstellen, wohingegen „NONE“ und „UNDEFINED“ auf einen nicht näher spezifizierten Zustand hinweisen. Es ist folglich ein Unterraum des Datensatzes zu wählen, in dem „PASS“ und ein „FAIL“ hinreichend oft vorhanden ist.

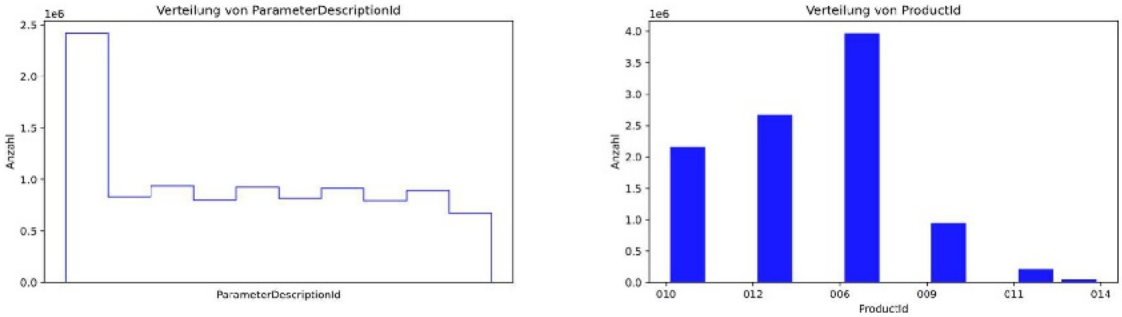

Abbildung 3-1 Histogramme zu den Attributen PDID und PID (eigene Darstellung)

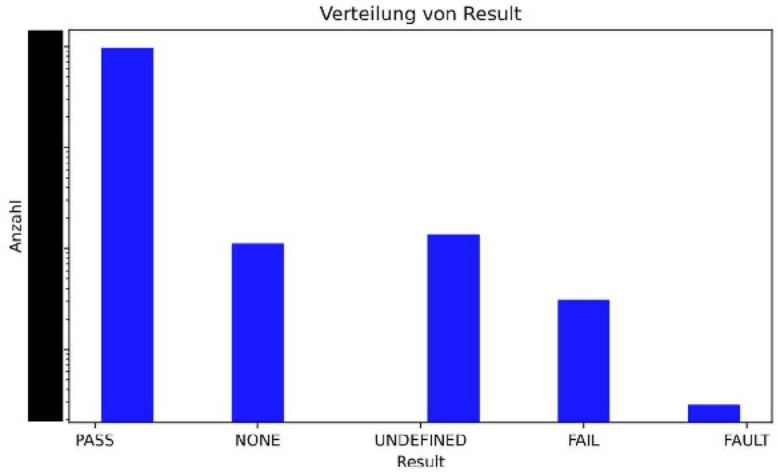

Abbildung 3-2: Histogramm zum Attribut Result in logarithmischer Darstellung (eigene Darstellung)

Der weitere Ansatz besteht darin, eine Menge an PDIDs zu finden, für die die zugehörigen Ausprägungen des Attributes Value einen numerischen Datentyp annehmen und sich auf einer kardinalen Skala befinden. In einem Datensatz aus numerischen und kardinalen Attributen sind die Bedingungen gegeben und mit den in 2.3 adressierten Verfahren Ausreißer zu detektieren. Aus der Menge der 76 verschiedenen PDIDs werden alle gefiltert, die im Attribute Result neben der Ausprägung „PASS“ keine Ausprägungen des Typs „FAIL“ enthalten. Die verbleibenden 38 PDIDs werden weiter nach dem vorliegenden Skalenniveau gefiltert. Hierfür wird ein weiteres Data Dictionary erstellt, dass alle PDIDs mit verschiedenen Results enthält.

Tabelle 3-2: Auszug des Data Dictionarys der PDIDs (eigene Darstellung)

Parameter-DescriptionId	Datentyp	Maximal Wert	Minimaler Wert	Mittelwert	Häufigster Wert	Skalenniveau	Ergebnisse
PDES000 00040	Integer	[redacted]			-	Kardinal	[redacted]
PDES000 00041	Object	-	-	-	NO ERROR	Nominal	
PDES000 00042	Integer	[redacted]			-	Kardinal	

Ein Auszug des Data Dictionary's ist in Tabelle 3-2 dargestellt und das vollständige Data Dictionary ist im Anhang in Tabelle 7-2 zu finden. Die gesuchten PDIDs zeichnen sich dadurch aus, dass der Datentyp nicht vom Typ String oder Object ist, in den Ergebnissen die Ausprägung „FAIL" enthalten ist und das Skalenniveau kardinal ist. Ein Beispiel für eine PDID, für die die Kriterien gelten, ist die PDID „040". Wie aus der Tabelle zu entnehmen ist, sind alle Bedingungen erfüllt. Nach dem Filtern nach diesen Kriterien sind an diesem Punkt dreizehn verschiedene PDIDs vorhanden, die für die Fragestellung infrage kommen.

Im Fall der zu prognostizierenden Werte handelt es sich um eine Reihe an Werten, die nach einem Zeitpunkt sortiert werden kann und die einen inhärenten Zusammenhang darstellen. Bei einer auf diese Weise dargestellte Menge von Werten handelt es sich um eine Zeitreihe. Eine Zeitreihe ist definiert als eine nach dem Zeitpunkt geordnete Reihe von Werten (Abedjan et al. 2019, S. 94). Im Rahmen einer Untersuchung einer Zeitreihe ist es sinnvoll, die Verteilung, Frequenz und Dichte der Werte zu betrachten (Abedjan et al. 2019, S. 94). Des Weiteren sind Muster wie Saisonalität und Regelmäßigkeiten im Verlauf interessante Betrachtungspunkte (Abedjan et al. 2019, S. 95).

Werden die verbleibenden PDIDs betrachtet, ist zu erkennen, dass verschiedene Muster in den Zeitreihen der zugehörigen Werte auftreten. In Abbildung 3-3 sind zwei verschiedene Typen von Mustern gegenübergestellt. Auf der linken Seite sind die Zeitreihen der PDIDs „040" und „092" abgebildet. Diese Zeitreihen zeigen eine Verteilung der Ausprägungen der Werte über einem Wertebereich, gepaart mit Abweichungen einzelner Werte aus diesem Bereich, auf. Demgegenüber stehen die Zeitreihen der der PDIDs „050" und „130". Hier ist zu erkennen, dass die Werte nur eine geringe Anzahl von konkreten Werteausprägungen annehmen. Es ist anzunehmen, dass eine Prognose mit einem Regressionsverfahren auf diesen konkreten Werten nicht zielführend ist. Für diese Fälle liegt eine Klassifikation nahe, welche wiederum nicht die Ansprüche an die Ausreißer Detektionsverfahren erfüllt, da hier nach einem kontinuierlichen Wertebereich gefragt ist.

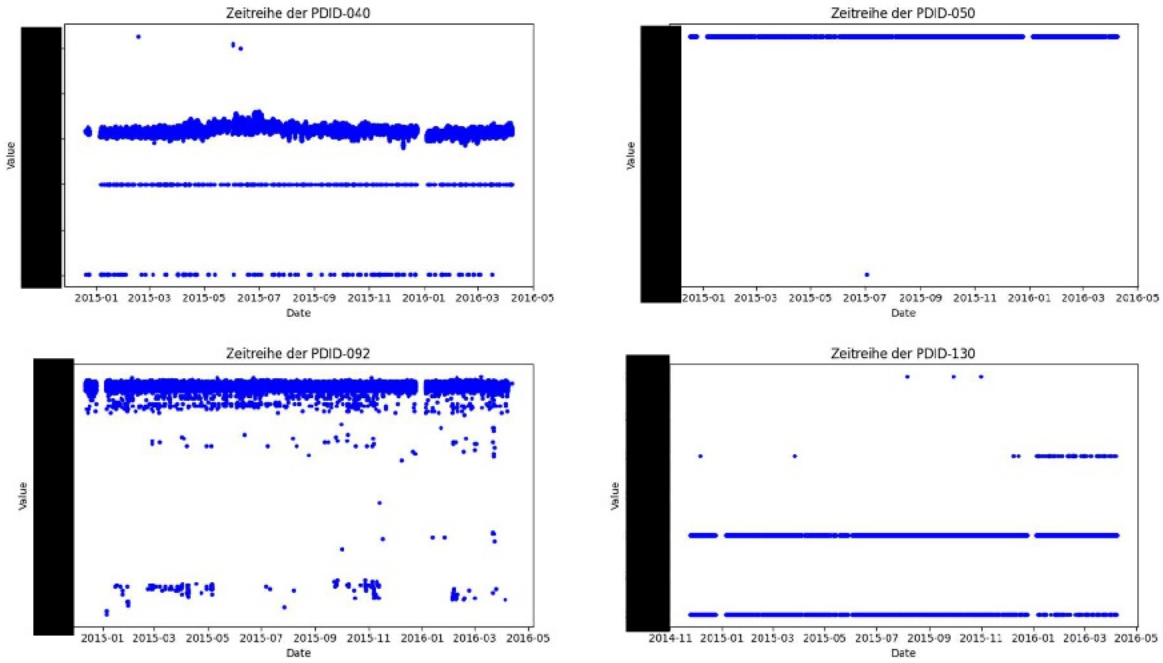

Abbildung 3-3: Zeitreihen der PDID 040, 050, 092 und 130 (eigene Darstellung)

Als Konsequenz aus dieser Beobachtung wird der verbleibende Datenraum auf die PDIDs weiter eingegrenzt, deren Zeitreihen auf einen stetigen Wertebereich schließen lassen und deren Ausprägungen nicht auf wenige konkrete Werte begrenzt sind. Nach dem Eingrenzen der PDIDs nach diesen Kriterien für eine Zeitreihenprognose sind die folgenden PDIDs für den finalen Datensatz geeignet: „040“; „042“; „091“; „092“; „110“; „111“.

Final liegen sechs verschiedene Zeitreihen vor, die jeweils für die Prognose der Werte verwendet werden und somit den weiter zu untersuchenden Unterraum des Datensatzes darstellen. Diese bilden die Grundlage, um die erschlossene Fragestellung zu bearbeiten. Die Fragestellung ist formuliert als Zeitreihenprognose der Ausprägungen des Attributes Value, für bestimmte Parameter, die durch die PDID festgelegt sind. Im Kontext dieser Fragestellung wird der Einfluss der Ausreißer innerhalb des Attributes ermittelt.

3.1.3 Datensätze für die Ausreißerentdeckung bestimmen

Basierend auf dem gefundenem Unterraum in den Daten muss konkretisiert werden, in welchen Datensätzen die Ausreißer identifiziert werden. Aufgrund der Auswahl des Unterraums des Datensatzes liegen nur noch Wertausprägungen vor, auf die statistische und distanzbasierte Verfahren anwendbar sind. Es lassen sich folglich Datensätze konstruieren, die für alle der beschriebenen Ausreißerkennungsverfahren verwendbar sind.

Initial lassen sich die Zeitreihen der PDIDs jeweils einzeln verwenden. In diesem Fall können statistische Verfahren auf die Ausprägungen des Attributes Value angewandt werden. Des Weiteren können für die dichte- und distanzbasierten Verfahren zusätzliche numerische Werte, wie beispielsweise die Zeitdifferenz zwischen den einzelnen Werten, hinzugezogen werden.

Außerdem kann an dieser Stelle eine Fallunterscheidung eingeführt werden. Es ist davon auszugehen, dass Fachpersonal, welches den Prozess kennt, weiß, dass einige der zu identifizierenden Ausreißer akausal sind. Dies bedeutet, dass diese Art von Abweichungen nicht detektiert werden muss, da sie im Voraus ausgeschlossen werden können. In Abbildung 3-4 ist die Zeitreihe der PDID 040 zweimal dargestellt. In der linken Abbildung sind alle Werte dargestellt und in der rechten Abbildung sind die Werte, die als Störung bzw. akausale Werte angesehen werden können, ausgeschlossen. Im Weiteren können also sowohl auf die vollständige

Zeitreihe als auch auf die reduzierte die verschiedenen Ausreißererkennungsverfahren angewandt werden.

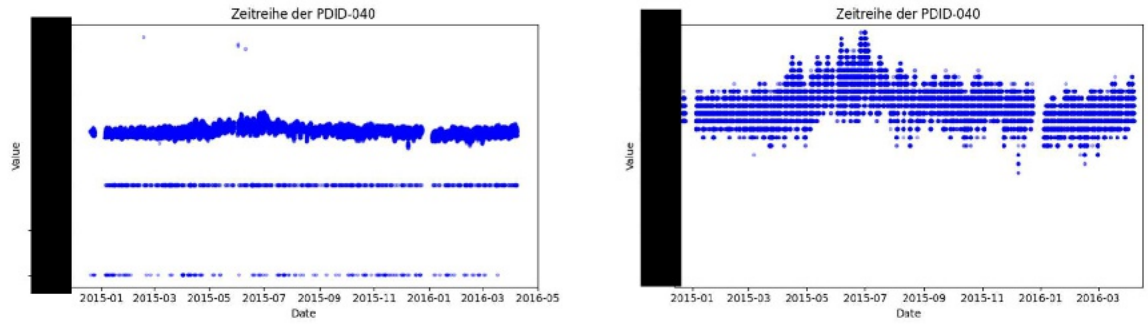

Abbildung 3-4: Zeitreihen der PDID 040: Vollständig, Ausschnitt (eigene Darstellung)

Neben der Betrachtung einzelner Zeitreihen besteht die Möglichkeit für einen ganzheitlichen Ansatz der Ausreißerdetektion. Die Ausprägungen des Attributes Value lassen sich jeweils einem konkreten Werkstück zuordnen. Hieraus lässt sich ein Zusammenhang der einzelnen Tupel herstellen, indem die Werte eines Tupels über das Identifikationsmerkmal der Werkstücke, die WpG, vereint werden. Im Kontext des Produktionssystems bedeutet dies, dass ein Werkstück verschiedene Bearbeitungsschritte und zugehörige Messungen durchläuft. Es lassen sich mehrere Messwerte einem konkreten Werkstück zuordnen. Die verschiedenen Werte der Messungen unterscheiden sich in der zugehörigen PDID. Auf diese Weise lässt sich ein Datensatz erzeugen, in dem die verschiedenen Messwerte, die zu einer WpG gehören, zusammengeführt und auf Basis der PDIDs verglichen werden. In Abbildung 3-5 ist ein Auszug eines solchen Datensatz für die PDID 040, 042 und 091 dargestellt. Jeder Punkt in dem gezeigten Diagramm enthält drei Messwerte eines Werkstückes. In Abbildung 3-6 ist der gleiche Würfel in den drei möglichen Projektionsebenen dargestellt. Mit Hilfe von distanz- und dichtebasierten Ausreißererkennungsverfahren kann in diesem Datensatz nach Ausreißern gesucht werden. Diese gefundenen Ausreißer stellen dann wiederum Werkstücke dar, deren Messungen zusammen eine Anomalie aufzeigen.

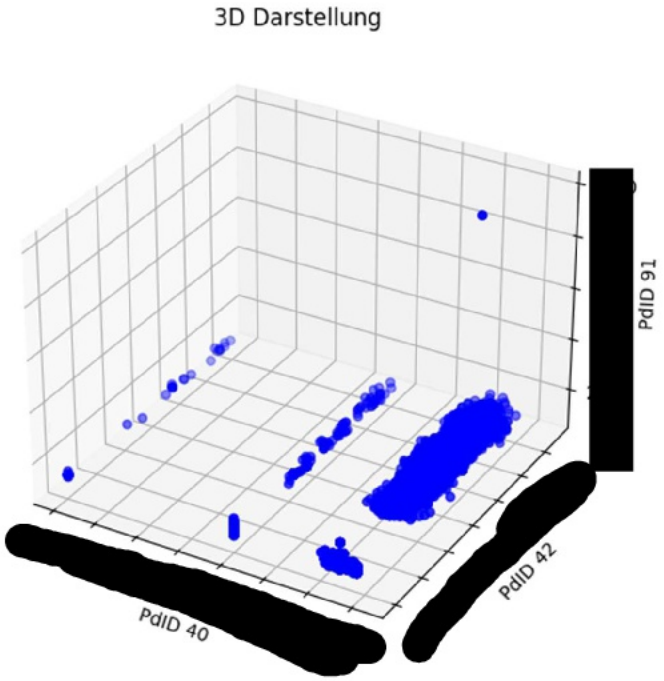

Abbildung 3-5: 3D-Visualisierung verschiedener PDIDs (eigene Darstellung)

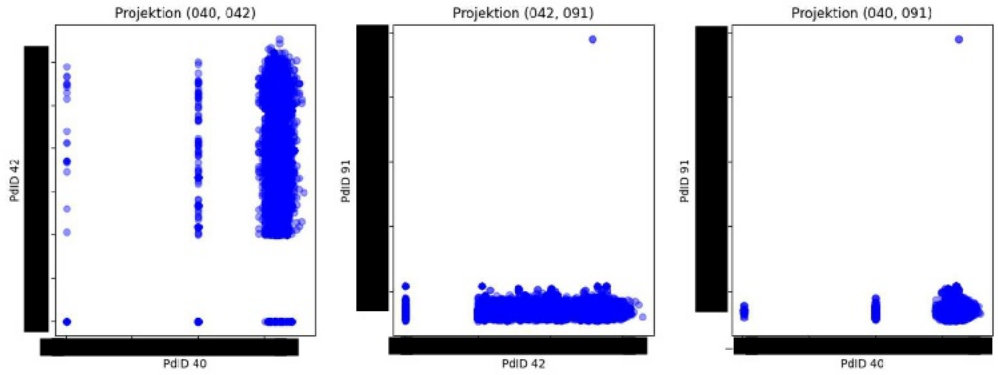

Abbildung 3-6: Projektionen verschiedener PDIDs (eigene Darstellung)

Der erstellte Datenraum enthält 1,8 Millionen Tupel, entsprechend viele verschiedene Werkstücke und je einen Wert für jede der sechs PDIDs.

Es folgt die Planung und Umsetzung eines geeigneten Modells zur Zeitreihenprognose. Der zu untersuchende Einfluss einer Detektion und Behebung von Ausreißer wird parallel dazu erarbeitet.

3.2 Datenvorverarbeitung und Umsetzung des Modells im Kontext der Ausreißererkennung

Im CRISP-DM Modell folgt auf das Data Understanding die Datenvorverarbeitung. Im Data und Business Understanding ist die zu bearbeitende Aufgabenstellung als Zeitreihenprognose numerischer Messerwerte festgelegt worden. Hieraus ergeben sich konkrete Anforderungen an die Datenvorerarbeitung. Diese Anforderungen an die Datenvorverarbeitung werden weiter durch das gewählte Prognosemodell spezifiziert. Aufgrund dieser Wechselwirkung aus Modellauswahl und Datenvorverarbeitung werden in diesem Abschnitt beide Themen simultan erarbeitet.

Allgemein lässt sich zur Modellauswahl festhalten, dass ein Modell gefragt ist, dass neben den numerischen Daten der Messwerte aus dem Value Attribut zu einer PDID auch die Attribute, des Produktionssystems miteinbeziehen kann. Heißt, es soll neben dem zeitlichen Verlauf des Attributes Value auch nach zusammenhängenden Mustern in den weiteren Attributen gesucht werden. Es liegt die Annahme zugrunde, dass die weiteren Attribute wie der Produkttyp einen Einfluss auf den gemessenen Wert haben. Prognosemodelle, die in der Lage sind, verschiedene Faktoren in einer Regression zu verarbeiten, sind unter den Machine-Learning-Modellen zu finden. Hier bieten sich Machine-Learning-Verfahren an, die auf Neuronale-Netze oder Entscheidungsbäume basieren. Mögliche Ausprägungen dieser Modelltypen sind: LSTMs, RNNs oder Random Forrests und Extreme Gradient Boosted Forrests (Garg und Singh 2021, S. 615; Moroff et al. 2021, S. 46). Diese Modelle stellen jeweils mögliche Verfahren für den hier vorliegenden Anwendungsfall dar. Als Entscheidungskriterium ist im Rahmen der durchzuführenden Experimente besonders die Trainingsdauer und die Genauigkeit relevant, da eine hohe Zahl von Experimenten mit jeweils neu trainierten Modellen möglich sein sollen. Auf Grund dieser Kriterien ist die Wahl auf den Extreme-Gradient-Boosted-Regression-Forrest kurz XGB gefallen (Garg und Singh 2021, S. 615, 2021, S. 614).

Das XGB-Modell nutzt Entscheidungsbäume, die in Form von Random Forrests aggregiert werden. Der Entscheidungsbaum ist ein Modell aus dem überwachten maschinellen Lernen, dessen Funktion es ist, Daten anhand von Entscheidungen in Teilmengen aufzuteilen. Die Entscheidungen basieren auf den Ausprägungen der bereitgestellten Daten und die finalen Teilmengen bilden die Klassen der Prognose (Ray 2019, S. 37; Garg und Singh 2021, S. 613).

Ein Random Forrest nutzt mehrere Entscheidungsbäume, die unter unterschiedlichen Startkriterien erzeugt werden und die über Mittelwertbildung zum Ergebnis des Random Forrest führen (Garg und Singh 2021, S. 613). Im Fall des Gradient Boostings wie im Fall des XGB-Modells werden die einzelnen Bäume des Random Forrests auf Basis der Residuen des vorherigen Baumes erzeugt, sodass eine schrittweise Optimierung der einzelnen Bäume erfolgt (Garg und Singh 2021, S. 613; Moroff et al. 2021, S. 42). Aus der Data Mining Fragstellung und dem gewählten Modell ergeben sich Anforderungen an die Form der Daten.

3.2.1 Attribute

Im ersten Schritt muss der gesamte Datensatz nach den verschiedenen PDIDs gefiltert werden. Die gefilterten Daten werden dann entsprechend des Zeitpunktes sortiert. Aus dem Attribut des Zeitpunktes können im zweiten Schritt die zusätzlichen Attribute Jahr und Monat extrahiert werden. Zusätzlich kann die Zeitdifferenz zwischen zwei Messpunkten über Differenzbildung ermittelt werden. Diese Schritte sind notwendig, um mit dem XGB-Modell Out-of-bound-Prognosen durchzuführen. Das ursprüngliche Attribut des Zeitpunktes kann in Random-Forrest-Modellen nur für Prognosen, die zeitlich innerhalb der Trainingsdaten liegen, verwendet werden. Da durch die Aufgabenstellung jedoch Prognosen von zukünftigen Werten gefordert sind, müssen für diese Modelle der Zeitpunkt in verschiedenen Attributen zerlegt werden. Mit diesen neuen konstruierten Attributen können dann zeitliche Zusammenhänge abgebildet werden, die nicht auf den Zeitraum der Trainingsdaten beschränkt sind.

Das Attribut Produkt-ID wird auf eine numerische Skala transformiert, da das Modell keine nicht numerischen Werte verarbeiten kann. Um die Verarbeitung zu optimieren, wird das erzeugte Attribut der Zeitdifferenz auf einen Wertebereich von 0 bis 1 skaliert.

Die finalen Daten werden dann in zwei Vektoren aufgeteilt, den Feature-Vektor und die Zielvariable. Der Feature-Vektor enthält die Attribute: WS, RoS, Res, PdID, Wochentag, Monat, Jahr, Zeitdifferenz (TimeDiff). Im Vektor der Zielvariablen ist das Zielattribut Value enthalten. Das Modell bildet dann den Zusammenhang zwischen den Features und der Zielvariablen ab. In Tabelle 3-3 ist ein Ausschnitt des verwendeten Featurevektors dargestellt. An diesem Punkt ist die Zeitdifferenz noch nicht skaliert. Das Attribut PDID ist kein Teil der Attribute zum Anlernen des Modells, da sich ein Modell immer nur auf die Zeitreihe einer einzigen PDID bezieht.

Tabelle 3-3: Ausschnitt des Featurevektors (eigene Darstellung)

Work-Sequence	Routing-Sequence	Result-Sequence	Product-Id	Weekday	Month	Year	TimeDiff
6	50	5	0	4	12	2014	0,00E+00
6	50	5	2	4	12	2014	3,3E-04
6	50	5	0	4	12	2014	4,23E+10
9	50	5	0	4	12	2014	2,80E-04
9	50	5	0	4	12	2014	1,89E-04

Um ein Modell anzulernen und im Anschluss testen zu können, wird der fertige Datensatz aufgeteilt. Die zeitlich betrachteten ersten 80% der Daten dienen zum Anlernen und werden im Weiteren als Trainingsdaten bezeichnet. Die aktuellen 20% der Daten sind die Testdaten und dienen dazu die Prognosegüte zu bestimmen. Mit dieser Einteilung kann die tatsächliche Fähigkeit des Modells, zukünftige Werte zu prognostizieren, bestimmt werden.

3.2.2 Parametertuning

Für die Durchführung von zielgerichteten Experimenten muss das gewählte Modell geeignet parametrisiert werden. Hierfür werden die zu parametrisierenden Parameter gewählt, welche in Tabelle 3-4 aufgelistet und beschrieben sind. Die Beschreibungen der Parameter basieren auf der Modelldokumentation von Pedregosa et al. (2011).

Tabelle 3-4: Parameter des XGB-Modells nach Pedregosa et al. (2011)

Parameter	Beschreibung	Gewählter Wertebereich	Ausprägung
Max Depth	Beschreibt die maximale Tiefe der einzelnen Bäume, welche sich aus der Anzahl der Knoten eines Astes ergibt.	3 – 5	3
Learning Rate	Einfluss des nächsten Baumes auf die Gesamtwertung	0,001 – 0,1	0,01
N-Estimators	Gibt die Anzahl an verwendeten Bäumen, Boosting-Stufen an	100 – 800	400

Um eine geeignete Kombination an Ausprägungen zu finden, wird der Raum der Parameterausprägungen systematisch abgesucht. Hierfür wird ein Grid-Search-Algorithmus verwendet. Dieser kann den Raum aus möglichen Ausprägungen durch Testen aller Kombinationen erschließen. Als Gütemaß wird der quadratische Fehler verwendet. Im Kontext der Zeitreihenprognose ist für die Kreuzvalidierung der Time-Series-Split gewählt, welcher durch zufälliges Ziehen von Datensätzen ermöglicht, in der Kreuzvalidierung die zeitlichen Zusammenhänge der Daten mitzuberücksichtigen. Im Detail ist dies nachzulesen in der Dokumentation der zugehörigen Bibliothek von Pedregosa et al. (2011).

3.2.3 Technische Evaluation der Prognose

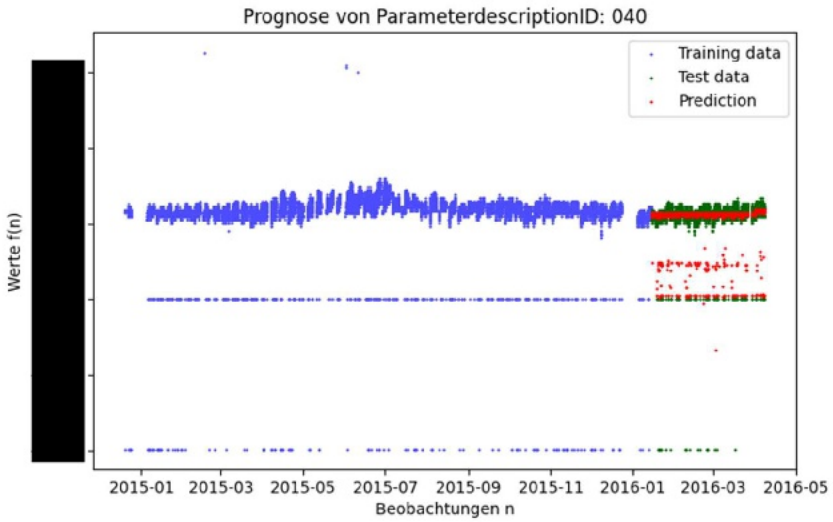

Abbildung 3-7: Prognose für PDID 040 (eigene Darstellung)

Für die Evaluation der Prognose werden die einzelnen Zeitreihen prognostiziert. Eine gewählte Prognose, dargestellt in Abbildung 3-7, zeigt, dass das Modell den Verlauf der Zeitreihe approximier und in der Lage ist bestimmte Muster für zukünftige Werte zu reproduzieren. Die hier gezeigten prognostizierten Werte zwischen 0 und [redacted] zeigen aber auch, dass die Mittelwertbildung des Random Forrests einen systematischen Fehler aufweist, da nicht eindeutig vorhergesagt wird, ob ein Wert die im Bereich über [redacted] oder bei 0 liegt und so der Mittelwert keinen der beiden Bereiche trifft. Es bleibt im Rahmen der Evaluation zu prüfen, wie die Erkennung von Ausreißern sich auf diese Art der Abweichungen auswirkt.

3.3 Implementierung der Ausreißererkennungsverfahren

Die Umsetzung der in Kapitel 2.3 beschriebenen Verfahren der Ausreißererkennung werden im Weiteren erläutert. Die Ausreißererkennung wird nur in den Trainingsdaten durchgeführt. Eine Ausnahme ist der Sonderfall des Datenraums. In der nachfolgenden Tabelle 3-5 sind die einzelnen Verfahren aufgelistet. Zusätzlich werden in der Tabelle die vom Verfahren genutzten Kennzahlen sowie die eingestellten Parameter beschrieben. Zur Vereinfachung werden die genutzten Daten, also die Daten, in denen nach Ausreißern gesucht wird, abgekürzt. Der Featurevektor ist mit X und die Zielvariable, das Attribut Value, mit Y beschrieben, gemäß den Konventionen der Scikit-Learn Dokumentation nach Pedregosa et al. (2011).

In den statistischen Verfahren ist der einzige einzustellende Parameter der Faktor der Standardabweichung, um den ein Wert vom Lagemaß abweichen darf. Dieser Parameter ist als Grenze angegeben. Im Fall des rollierenden Lagemaßes wird zusätzlich die Fenstergröße festgelegt, um die das Lagemaßes bestimmt wird. Für die distanzbasierten Verfahren wird die Anzahl der umliegenden Werte festgelegt, bezeichnet als n oder k, sowie die minimale Losgröße für das DBSC-Verfahren. Im Fall des DBSC-Verfahrens kann auch der maximale Abstand festgelegt werden, innerhalb dessen ein Punkt zu einem Cluster gehört. Dieser Abstand wird als Epsilon bezeichnet. Die Beschreibungen der Parameter der distanzbasierten Verfahren basieren auf der Dokumentation der verwendeten Scikit-Learn-Bibliothek von Pedregosa et al. (2011).

Tabelle 3-5: Implementierung der Ausreißererkennungsverfahren (eigene Darstellung)

Verfahren	Kennzahlen	Parametrisierung	Genutzte Daten
Abweichung vom Mittelwert	Mittelwert, Standardabweichung, Grenze	Grenze = 2	Y
Abweichung vom Median	Mittelwert, Standardabweichung, Grenze	Grenze = 2	Y
Abweichung vom rollierenden Mittelwert	Mittelwert, Standardabweichung, Grenze, Zeitfenster	Grenze = 2; Zeitfenster = [redacted]	Y
Abweichung vom rollierenden Median	Mittelwert, Standardabweichung, Grenze, Zeitfenster	Grenze = 2, Zeitfenster = [redacted]	Y
LOF	N-Neighbors	N = 20	X und Y
DBSC	Epsilon (Eps), Min_Samples (min)	Eps = 0.5, Min = 10	X und Y
KNN	K, Schwellwert	K = 5, Schwellwert = 2	X und Y
LOF im Datenraum	N-Neighbors	N = 20	X (für alle PDID) und Y

Für den Datenraum der Werkstücke wird das LOF-Verfahren zur Identifikation der anomalen Werkstücke verwendet.

Die angewandten Verfahren, um Ausreißer zu beheben, sind folgende Imputationsverfahren: Das Löschen der Werte sowie das Ersetzen der Werte mit statistischen Kennzahlen. Als statistische Kennzahlen zum Ersetzen von Ausreißern werden der Median und der arithmetische Mittelwert genutzt.

Auf der Grundlage der durchgeführten Datenvorverarbeitung und dem gewählten Modell können im Weiteren einzelne Prognosen für die verschiedenen Zeitreihen in Kombination mit den Ausreißererkennungsverfahren durchgeführt werden. Die resultierenden Prognoseergebnisse werden im nächsten Kapitel untersucht.

4 Evaluation der Ausreißeruntersuchung

In diesem Kapitel folgen die Auswertungen der Zeitreihenprognosen für die vorverarbeiteten Datensätze, die bereits in Kapitel 3.2 beschrieben werden. Darüber hinaus werden Schlussfolgerungen bezüglich der untersuchten Ausreißererkennungsverfahren gezogen. Zur Evaluation der Vorhersagegenauigkeit werden als Kennzahlen der MAE und der RMSE betrachtet, die in Kapitel 2.1.2 erläutert werden. Außerdem wird die Anzahl entdeckter Ausreißer betrachtet, um die Ausreißererkennungsverfahren auch in dieser Hinsicht miteinander zu vergleichen.

4.1 Evaluation der Ausreißererkennungsverfahren für die vorverarbeiteten Datensätze (PDIDs)

Wie bereits in Kapitel 3.3 erwähnt, werden die sieben ausgewählten Ausreißererkennungsverfahren und die drei Imputationsverfahren auf alle sechs ausgewählten PDIDs angewendet. Damit ergibt sich ein Vollfaktorplan, der in Abbildung 4-1 visualisiert ist. Insgesamt ergeben sich 126 Zeitreihenprognosen, die im Folgenden auf die wesentlichen Ergebnisse reduziert miteinander verglichen werden, um Rückschlüsse bezüglich der Güte der Ausreißererkennung mithilfe der verschiedenen Verfahren ziehen zu können.

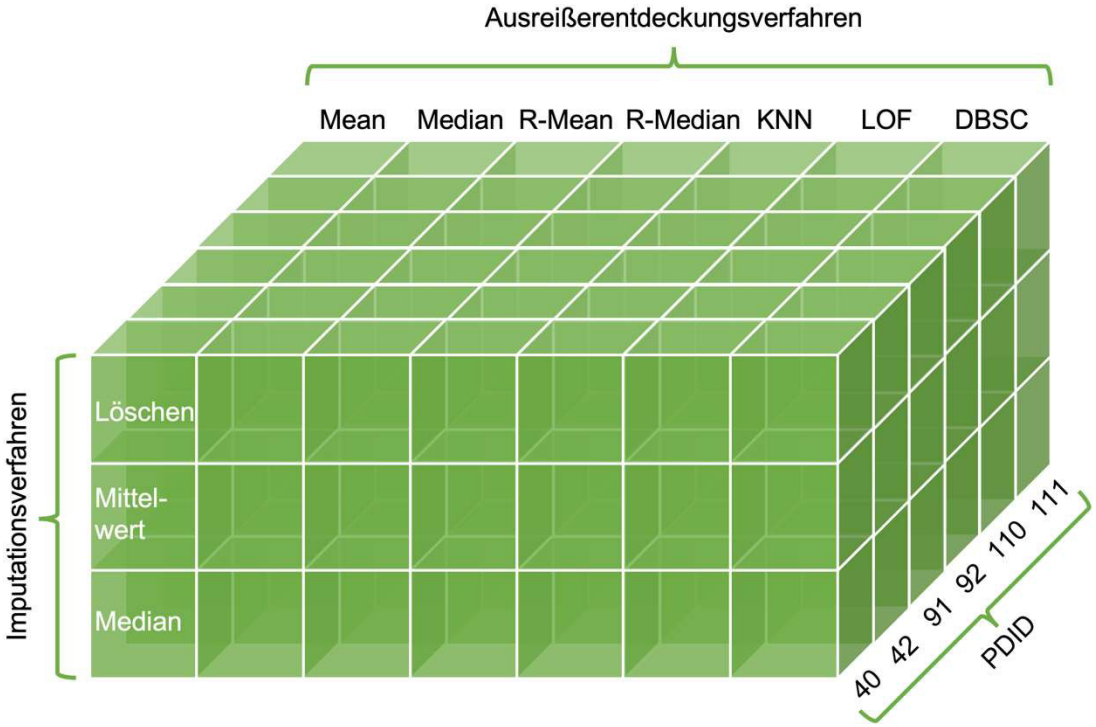

Abbildung 4-1: Versuchsplan zur Evaluation der Ausreißererkennungsverfahren (eigene Darstellung)

4.1.1 Vergleich der Anzahl gefundener Ausreißer je Ausreißererkennungsverfahren

Bei der Betrachtung der Anzahl gefundener Ausreißer wird schnell ersichtlich, dass abhängig vom Datensatz unterschiedliche Verfahren zu der höchsten Anzahl an Ausreißern führen. Bei den Datensätzen der PDID 40, 92, 110 und 111 werden durch das LOF-Verfahren mit Abstand am meisten Ausreißer gefunden, gefolgt vom DBSC-Verfahren. Für die PDID 40 ist dieser Sachverhalt in Abbildung 4-2 dargestellt. Die Datensätze der PDIDs 92, 110 und 111 weisen eine ähnliche Verteilung der gefundenen Ausreißer wie der Datensatz der PDID 40 auf und sind im Anhang (Abbildung 7-2 ff.) hinterlegt.

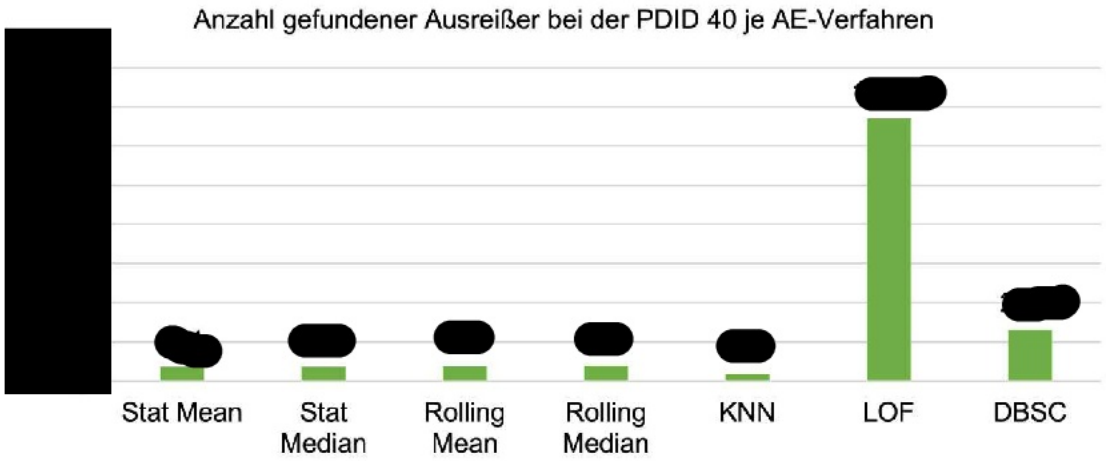

Abbildung 4-2: Anzahl gefundener Ausreißer je AE-Verfahren für PDID 40 (eigene Darstellung)

Bei den Datensätzen der PDID 42 und 91 hingegen werden am meisten Ausreißer über das DBSC-Verfahren entdeckt, gefolgt vom statistischen Verfahren „gleitender Median“. Die Verteilung der Anzahl gefundener Ausreißer für die PDID 42 ist in Abbildung 4-3 zu erkennen. Die Verteilung der Anzahl gefundener Fehler je Ausreißererkennungsverfahren der PDID 91 sieht ähnlich aus und ist daher im Anhang (Abbildung 7-1) hinterlegt.

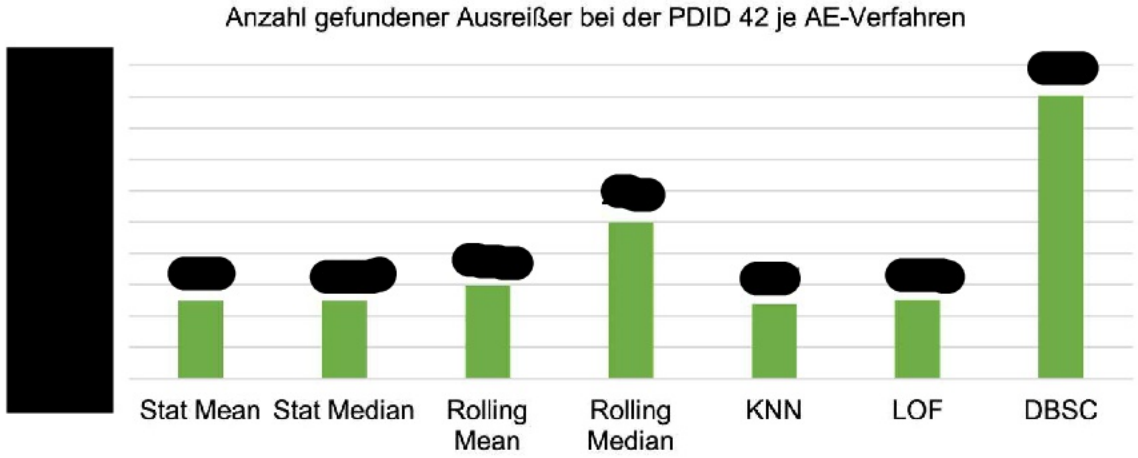

Abbildung 4-3: Anzahl gefundener Ausreißer je AE-Verfahren für PDID 40 (eigene Darstellung)

Anhand dieser Ergebnisse lässt sich demnach kein AE-Verfahren identifizieren, das grundsätzlich zu einer hohen oder niedrigen Anzahl entdeckter Ausreißer führt. Betrachtet man diesen Sachverhalt auf höherer Hierarchieebene, d.h. vergleicht man die Anzahl gefundener Ausreißer bei statistisch basierten und distanzbasierten AE-Verfahren, so lässt sich auch hier keine allgemeine Aussage bezüglich der Anzahl entdeckter Ausreißer treffen. Es ist in diesem Fall stark – und ggf. sogar einzig – abhängig vom gewählten Datensatz, welches Verfahren zu einer hohen Anzahl entdeckter Ausreißer führt.

4.1.2 Vergleich der Fehlerkennzahlen je Ausreißererkennungsverfahren

Da der RMSE als Fehlerkennzahl sensibler gegenüber Ausreißern ist als der MAE, wird dieser nachfolgend primär betrachtet. Die Ergebnisse bezüglich des MAE sind der Vollständigkeit halber im Anhang (Tabelle 7-3 ff.) aufgeführt, lassen jedoch dieselben Rückschlüsse zu wie die Betrachtung des RMSE als Performancekennzahl.

Zunächst fällt auf, dass der RMSE-Wert bei fünf der sechs betrachteten Datensätze für die drei Imputationsverfahren nahezu identisch ist (siehe Anhang, Tabelle 7-9 ff.). Lediglich bei dem Datensatz der PDID 40 ergeben sich für die AE-Verfahren KNN, LOF und DBSC mit unterschiedlichen Imputationsverfahren abweichende Werte, wie in Abbildung 4-4 zu sehen ist. Da die unterschiedlichen Imputationsverfahren also in 39 von 42 Fällen ($6\,Datensätze * 7\,Verfahren$) zu (nahezu) denselben Ergebnissen führen, wird zur Ergebnisdarstellung und Auswertung nachfolgend nur das Imputationsverfahren „Löschen“ betrachtet.

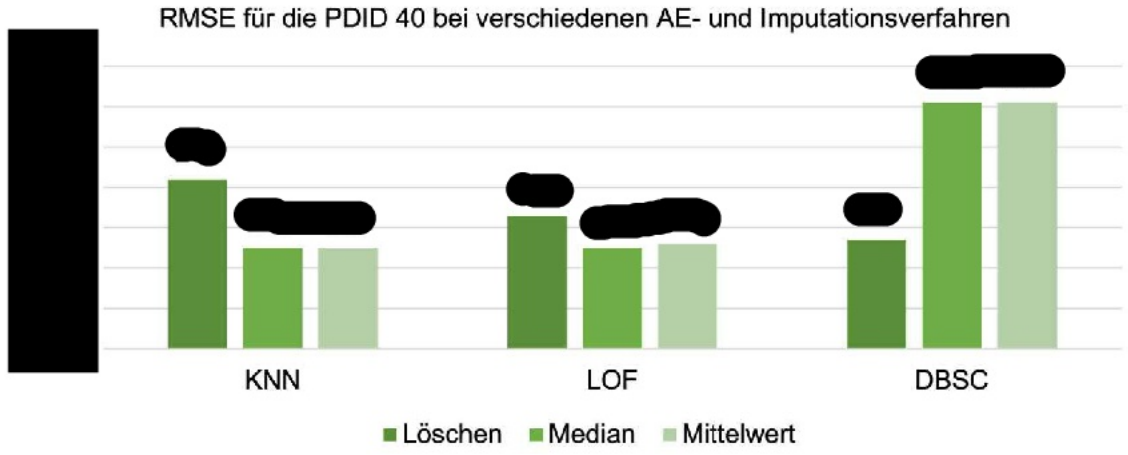

Abbildung 4-4: RMSE für die PDID 40 bei verschiedenen AE- und Imputationsverfahren (eigene Darstellung)

Auffallend ist ebenfalls, dass sich bei den Datensätzen der PDID 91 und 92 für alle AE-Verfahren dieselben RMSE-Werte nach der Imputation ergeben (siehe Anhang, Tabelle 7-11 und Tabelle 7-12). Dies liegt in den Werten der Zeitreihe begründet. Die Werte der PDID 92 befinden sich beispielsweise im Bereich von [redacted] (siehe Anhang, Abbildung 7-5). Vereinzelt befinden sich Datenpunkte auch im Bereich um den Wert [redacted] Da die Werte so nahe beieinander liegen, hat die Imputation von Ausreißern nur einen sehr geringen Einfluss auf die Vorhersagegenauigkeit. Dies wird deutlich, wenn man die Formel zur Berechnung des RMSE erneut betrachtet:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Die vorhandenen Datenpunkte weichen maximal um den Wert [redacted] voneinander ab. Geht man davon aus, dass das Modell die Werte nicht außerhalb des angegebenen Wertebereichs vorhersagt, so kann auch die Differenz zwischen dem prognostizierten Wert y_i und dem wahren Wert $\hat{y}_i$ maximal [redacted] betragen. Quadriert ergibt sich hier ein Wert von [redacted] Dies ist also der maximale Wert des RMSE, der sich für diese Zeitreihe – unter den getroffenen Annahmen bezüglich des Modells – ergeben kann. Es ist also wenig verwunderlich, dass die Werte für den RMSE für unterschiedliche AE-Verfahren so nahe beieinander liegen, wenn der maximale Wert des RMSE ohnehin sehr niedrig ist.

Über die Betrachtung des RMSE lässt sich im Allgemeinen schlussfolgern, wie präzise das Modell die Zeitreihe prognostizieren kann. Je kleiner dieser Wert ist, desto besser ist die Vorhersage (vgl. Kapitel 2.1.2). Die Erwartungshaltung vor der Auswertung ist, dass die Vorhersage besser wird, je mehr Ausreißer entfernt werden, da Ausreißer für gewöhnlich nicht systematisch auftreten und daher schlechter vom Modell prognostiziert werden können.

Betrachtet man jedoch den RMSE für die angewandten AE-Verfahren und vergleicht diesen mit dem Ausgangswert, so wird diese Erwartung nicht erfüllt. Der Ausgangswert ist der RMSE für eine Prognose ohne vorhergehende Ausreißererkennung, d.h. alle Ausgangswerte bleiben

im Datensatz erhalten. In Abbildung 4-5 ist erkennbar, dass die Fehlerkennzahl für den Datensatz der PDID 40 mit Anwendung der AE-Verfahren steigt und die Prognose nach der Entfernung der Ausreißer schlechter wird. Der Ausgangswert ist ebenfalls für die PDID 42, 110 und 111 (nahezu) der beste Wert. Im Anhang sind die Diagramme für diese Datensätze hinterlegt (Abbildung 7-6 ff.).

Wie in Abbildung 4-5 ebenfalls erkennbar ist, ist der RMSE beispielsweise nach einer Ausreißererkennung und -beseitigung mittels DBSC-Verfahren niedriger als nach einer Ausreißerbehandlung mittels Medians. Daraus lässt sich schließen, dass das AE-Verfahren durchaus einen Einfluss auf die Güte der Prognose hat, diese jedoch im Fall der ausgewählten Datensätze verschlechtert.

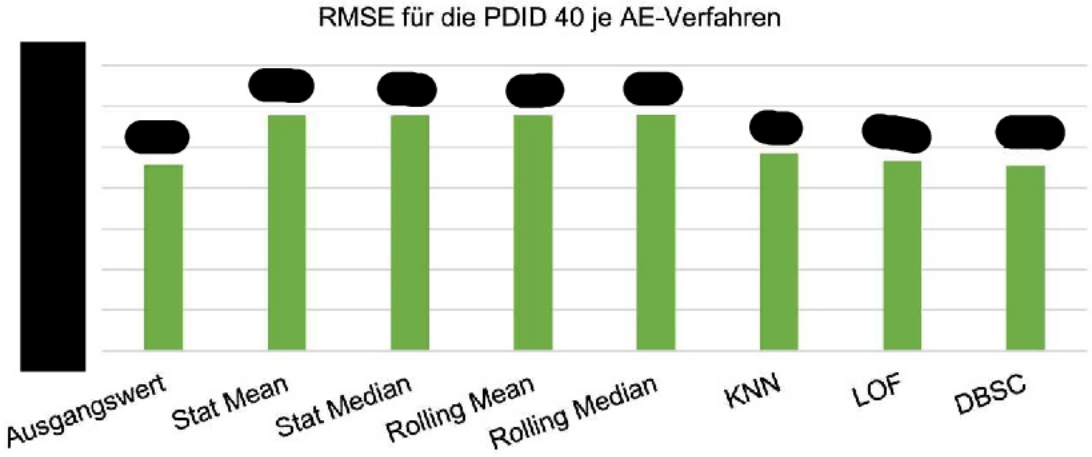

Abbildung 4-5: RMSE für den Datensatz der PDID 40 je AE-Verfahren (eigene Darstellung)

An dieser Stelle stellt sich darüber hinaus die Frage, ob gegebenenfalls ein Zusammenhang zwischen der Vorhersagegenauigkeit (RMSE) und der Anzahl gefundener Ausreißer besteht. Berechnet man für diesen Sachverhalt den Pearson-Korrelationskoeffizienten, so ergeben sich für die Datensätze jedoch lediglich Werte zwischen [redacted] und [redacted] (siehe Anhang, Tabelle 7-15 ff.). Es kann also keine starke lineare Korrelation zwischen der Anzahl der entdeckten Ausreißer und dem RMSE festgestellt werden. Auch wenn bei der Analyse der Ausgangswert, d.h. der RMSE für 0 Ausreißer, ausgeschlossen wird, ergibt sich kein aussagekräftiger Korrelationskoeffizient (siehe ebenfalls Tabelle 7-15 ff.). Eine höhere Anzahl entdeckter Ausreißer führt also nicht zwangsläufig zu einer Verbesserung der Prognose.

4.1.3 Entwicklung einer Fallunterscheidung für den Datensatz der PDID 40 zur optimierten Ausreißerentdeckung

Die Tatsache, dass die Vorhersagegenauigkeit für den Datensatz vor der Entdeckung und Imputation der Ausreißer am höchsten ist, lässt darauf schließen, dass das Modell auch Ausreißer vorhersagen kann. Es herrscht also möglicherweise eine Systematik im Auftreten von Ausreißern vor, die gegebenenfalls in der Datenvorverarbeitung schon erkannt und beseitigt werden kann. Hierfür wird nachfolgend eine Fallunterscheidung entwickelt, um die Entdeckung von Ausreißern und damit die Vorhersagegenauigkeit zu optimieren. Da es sich hierbei um einen manuellen Ansatz der Vorverarbeitung handelt, wird dieser vorerst nur auf den Datensatz der PDID 40 angewendet. Die Ergebnisse bezüglich dieser Fallunterscheidung werden nachfolgend mit „PDID 40 – Fall 2“ bezeichnet.

Hierfür wird die Zeitreihe der PDID 40 nach der Datenvorverarbeitung (siehe Kapitel 3.2) und vor der Anwendung von AE-Verfahren betrachtet. Die Zeitreihe ist in Abbildung 4-6 dargestellt. Zu sehen sind Schwankungen der Werte im Bereich um den Wert [redacted]. Darüber hinaus treten Datenpunkte bei den Werten [redacted] auf. Vereinzelt treten auch Datenpunkte im Bereich um den Wert [redacted] auf.

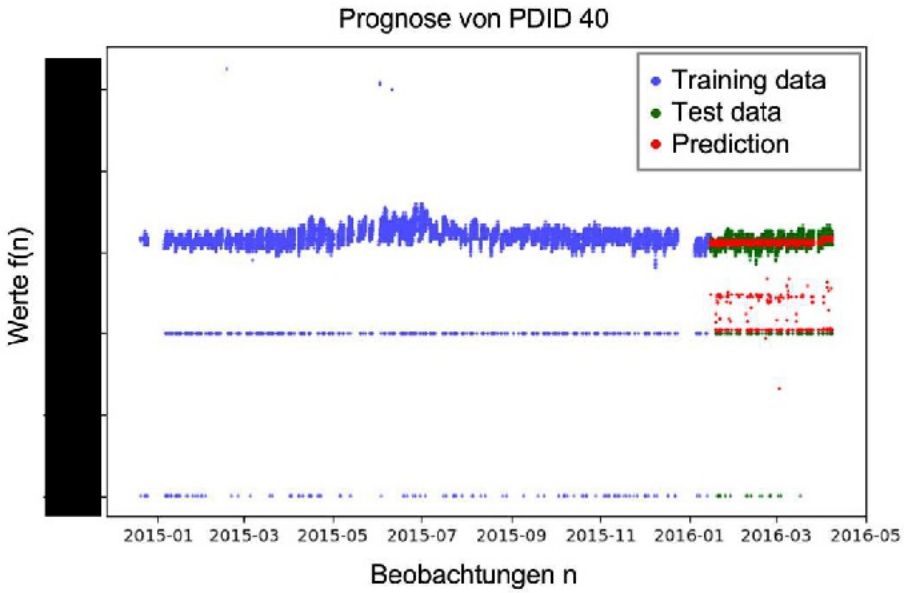

Abbildung 4-6: Zeitreihenprognose für die PDID 40 nach der Datenvorverarbeitung und vor Anwendung von AE-Verfahren (eigene Darstellung)

An dieser Stelle lässt sich die Vermutung aufstellen, dass diese Werte gegebenenfalls auf einen abweichenden Betrieb hinweisen, bei dem beispielsweise eine Anlagenstörung vorliegt. Auf Grundlage dieser Annahme werden die Werte im Rahmen der Datenvorverarbeitung entfernt. Die These hierbei ist, dass die Datenvorverarbeitung mithilfe von Prozesswissen optimiert werden kann. Werte, die auch ohne maschinelle Verfahren der Ausreißererkennung im Vorfeld als Ausreißer oder Anomalien bekannt sind, sollten vor Anwendung der AE-Verfahren ausgeschlossen werden. So können Ausreißer, die weniger offensichtlich und nicht manuell auffindbar sind, besser mithilfe der maschinellen Verfahren identifiziert werden.

In diesem Fall wird davon ausgegangen, dass die Werte [redacted] unternehmensintern für eine Anlagenstörung, einen Ausfall o.ä. stehen und damit einen vorab bekannten Ausreißer darstellen. Die betroffenen visuell gefundenen Datenpunkte werden demnach im Rahmen der Datenvorverarbeitung imputiert. Anschließend werden auf den neu entstandenen Datensatz die bereits untersuchten AE-Verfahren angewendet.

Für die Auswertung werden die Ergebnisse des Datensatzes für die PDID 40 und die Ergebnisse des Datensatzes „PDID 40 – Fall 2“ miteinander verglichen, um festzustellen, ob eine Vorverarbeitung mit Prozesswissen zu einer Verbesserung der Ergebnisse führt.

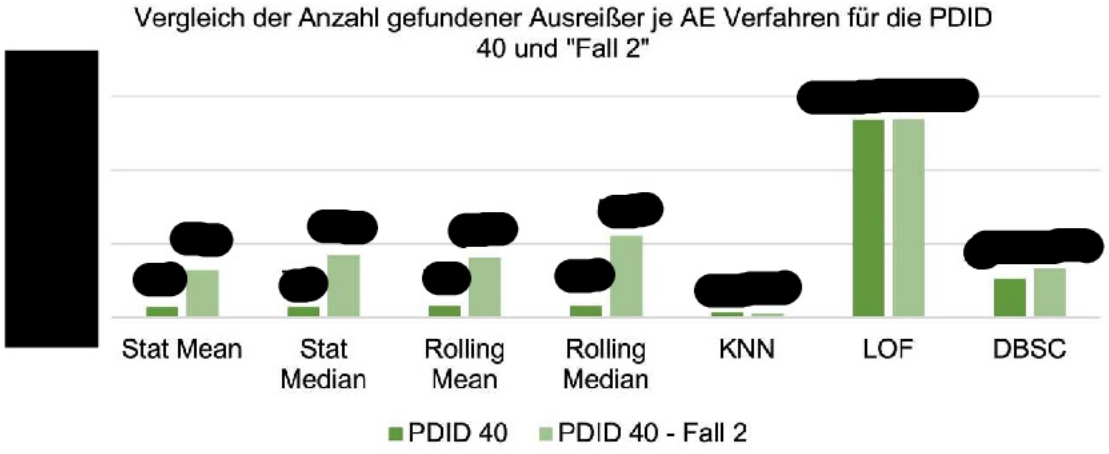

Abbildung 4-7: Entwicklung der Anzahl gefundener Ausreißer für die PDID 40 (eigene Darstellung)

Wie in Abbildung 4-7 zu sehen ist, ist im „Fall 2“ die Anzahl entdeckter Ausreißer für die statistisch basierten Verfahren deutlich gestiegen. Dies liegt vermutlich daran, dass sich das Toleranzband für Ausreißer verringert hat und hierdurch mehr Datenpunkte als Ausreißer erkannt werden. Bei den distanzbasierten Verfahren ist die Abweichung der Anzahl entdeckter Ausreißer weniger auffällig.

Von größerem Interesse ist jedoch, ob die optimierte Datenvorverarbeitung auch zu einer genaueren Prognose der Zeitreihe führt und ob der Einfluss der Ausreißerentdeckung für diesen Fall größer bzw. positiver ist. In Abbildung 4-8 ist die Entwicklung des RMSE für die Fallunterscheidung dargestellt. Der RMSE ist für den „Fall 2“ für jedes AE-Verfahren deutlich gefallen. Auch der Ausgangswert ohne Anwendung maschineller AE-Verfahren ist deutlich niedriger. Die Vorhersagegenauigkeit steigt also durch die zusätzliche Vorverarbeitung. Darüber hinaus ist eine weitere Verbesserung der Prognosegenauigkeit durch Anwendung maschineller AE-Verfahren in diesem Fall feststellbar. Die beste Vorhersagegenauigkeit wird demnach für den „Fall 2“ nach Anwendung des Mittelwert-Verfahrens erreicht mit einem RMSE von [redacted].

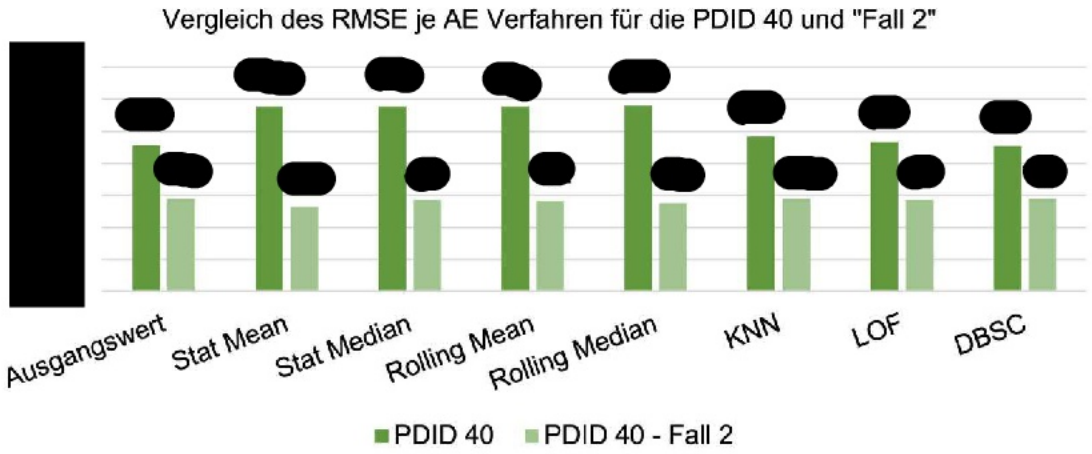

Abbildung 4-8: Entwicklung des RMSE je AE-Verfahren für die PDID 40 (eigene Darstellung)

Eine Erkenntnis, die aus dieser Fallunterscheidung folgt, ist also, dass sich die Datenvorverarbeitung mithilfe von Hintergrundwissen zu den Produktionsprozessen optimieren lässt und dadurch auch zu besseren Prognoseergebnissen führt. Es ist demnach wenig zielführend, AE-Verfahren „blind“ auf einen Datensatz anzuwenden. Bei der Behandlung von Ausreißern sollte immer Prozesswissen herangezogen werden und geprüft werden, ob die Imputation der entdeckten Ausreißer sinnvoll ist und zu einer Verbesserung der ML-Aufgabe führt.

Das hier beschriebene und angewendete Vorgehen für den Datensatz der PDID 40 kann theoretisch auch für die weiteren Datensätze durchgeführt werden. Aus Zeitgründen ließ sich dies jedoch im Rahmen des Fachlabors nicht umsetzen und ist ein offener Punkt für zukünftige Projekte in dieser Fachrichtung.

4.2 Evaluation der Ausreißererkennungsverfahren für den konstruierten Datenraum

Neben der Anwendung der AE-Verfahren auf die Datensätze der ausgewählten PDIDs wird, wie in Kapitel 3.2 beschrieben, auch die Ausreißererkennung im Datenraum untersucht. Der Datenraum ist ein Ansatz zur ganzheitlichen Betrachtung von Ausreißern im Prozess. Dabei sollen nicht einzelne Werte der unterschiedlichen PDIDs als Ausreißer erkannt und imputiert werden. Vielmehr sollen ganze Werkstücke – beschrieben durch die WorkpieceGuID – als Ausreißer identifiziert werden. Hierfür werden die Datensätze der gewählten PDIDs zu einem mehrdimensionalen Datenraum kombiniert. Wie auch in Kapitel 4.1.3 handelt es sich hierbei

um einen explorativen Ansatz, daher wird vorerst nur ein AE-Verfahren für die Untersuchung gewählt und auch nur ein Imputationsverfahren umgesetzt. Da es sich um mehrdimensionale Daten handelt, sind die statistischen Verfahren zur Ausreißerentdeckung ungeeignet. Die Wahl fällt daher auf das clusterbasierte LOF-Verfahren. Zur Imputation wird – wie auch in den vorhergehenden Analysen – das Verfahren „Löschen“ gewählt.

4.2.1 Anzahl gefundener Ausreißer und Einordnung in den Datenraum

Zunächst wird untersucht, wie viele Datenpunkte als Ausreißer identifiziert werden und wie diese visuell in den Datenraum eingeordnet werden können. Insgesamt werden ca. [redacted] Datenpunkte als Ausreißer pro Zeitreihe über das LOF-Verfahren identifiziert und damit für die Zeitreihenprognose gelöscht. In Abbildung 4-9 sind die Ausreißer zusammen mit den nicht als Ausreißer identifizierten Datenpunkten abgebildet. Die Ausreißer sind dabei blau markiert. Entgegen der Erwartung lässt sich hierbei visuell keine Punktwolke, die lediglich aus Ausreißern besteht, erkennen. Vielmehr befinden sich Ausreißer und Nicht-Ausreißer nahe beieinander, wobei diese Aussage nur auf optische Beobachtungen gestützt getroffen werden kann. Die Entdeckung der Ausreißer ist bei diesem ganzheitlichen Ansatz demnach nicht so offensichtlich möglich wie bei eindimensionalen Datensätzen, bei denen sich Ausreißer graphisch leicht erkennen lassen (vgl. z.B. Kapitel 2.3.1).

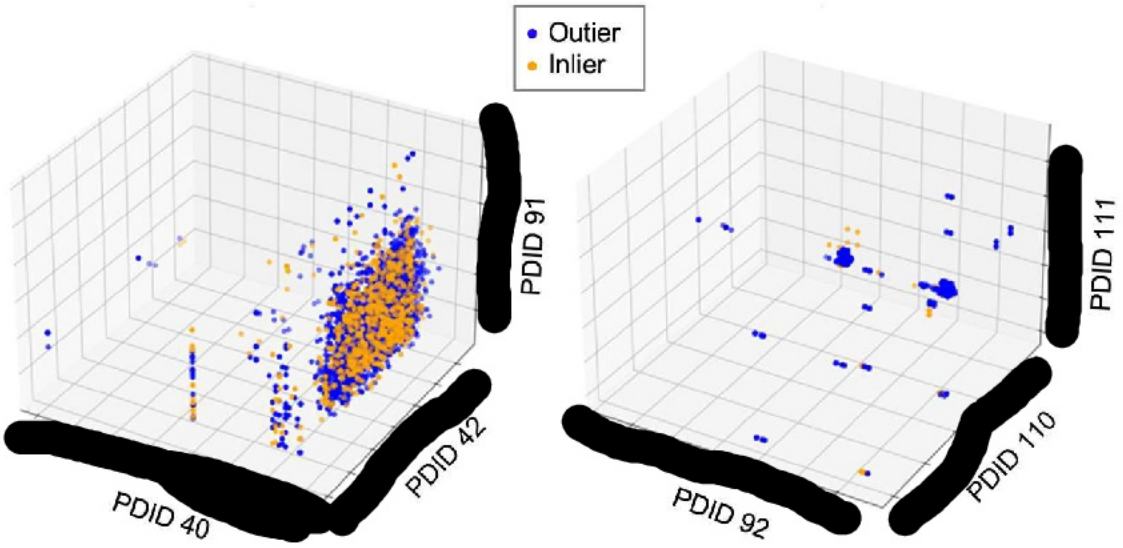

Abbildung 4-9: Visualisierung des Datenraums und Markierung der Ausreißer in blau, links für die PDIDs 40, 42 und 91, rechts für die PDIDs 92, 110 und 111 (eigene Darstellung)

4.2.2 Auswertung der Fehlerkennzahlen für die Ausreißerentdeckung im Datenraum

Auch an dieser Stelle wird der RMSE für die Zeitreihenprognose betrachtet, um Schlussfolgerungen bezüglich der Ausreißerentdeckung im Datenraum ziehen zu können. Hierbei stellt sich ebenfalls die Frage, ob die Anwendung von AE-Verfahren im Datenraum als Teil der Datenvorverarbeitung zu einer optimierten Prognose führt.

In diesem Zusammenhang werden der RMSE für den Ausgangsdatensatz und der RMSE nach der Ausreißerentdeckung und -imputation im Datenraum miteinander verglichen. Der RMSE des Ausgangsdatensatzes wird bereits in Kapitel 4.1.2 vorgestellt. In Abbildung 4-10 ist diese Entwicklung des RMSE graphisch dargestellt. Der RMSE verbessert sich wesentlich für die PDID 40, 110 und 111. Für die PDID 42, 91 und 92 hingegen bleibt die Vorhersagegenauigkeit unverändert, wobei eine Verbesserung der Prognosegüte bei der PDID 92 ohnehin nicht möglich ist.

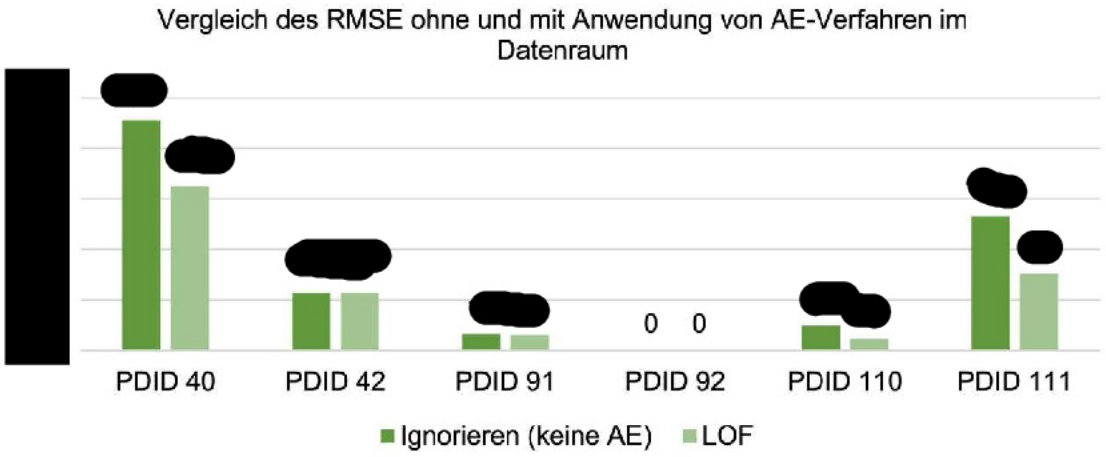

Abbildung 4-10: Entwicklung des RMSE je PDID nach der AE-Behandlung im Datenraum (eigene Darstellung)

Abschließend werden die Ergebnisse für diesen experimentellen, ganzheitlichen Ansatz wiederum mit den Ergebnissen der Fallunterscheidung in Kapitel 4.1.3 verglichen. Da die Fallunterscheidung lediglich für die PDID 40 durchgeführt wurde, kann die Gegenüberstellung der Ergebnisse an dieser Stelle auch nur für die PDID 40 erfolgen. Wie in Abbildung 4-10 zu sehen ist, beträgt der Ausgangswert des RMSE [redacted]. Nach Anwendung der AE-Verfahren im Datenraum sinkt dieser Wert auf [redacted] Bei der Fallunterscheidung ergibt sich ein RMSE von [redacted] Demnach ist die Vorhersagegenauigkeit nach einer zusätzlichen Ausreißererkennung mit Prozesswissen am besten. Der ganzheitliche Ansatz führt jedoch auch zu einer starken Verbesserung der Prognosegüte.

Insgesamt lässt sich also schlussfolgern, dass die beiden experimentellen Ansätze dieser Arbeit zur Optimierung der Ergebnisse der Machine-Learning-Aufgabe führen. Daraus lässt sich schließen, dass die Ausreißererkennung in diesen Fällen zu einer verbesserten Datenvorverarbeitung beitragen, die essenziell für das erfolgreiche Umsetzen von ML-Aufgaben ist.

4.3 Diskussion und Fazit

In den vorhergehenden Unterkapiteln wurden die Ergebnisse der Arbeit mit Blick auf die Aufgabenstellung zusammengefasst und ausgewertet. Ziel der Arbeit ist es, herauszufinden, welches Ausreißerbehandlungsverfahren sich am besten für die festgelegte Anwendung eignet. Die Anwendung besteht darin, Zeitreihenprognosen aufgrund der Values ausgewählter PDIDs zu erstellen.

Insgesamt lässt sich festhalten, dass hierbei kein AE-Verfahren identifiziert werden kann, das universell „am besten“ funktioniert. Vielmehr ist es stark vom betrachteten Datensatz abhängig, welches Verfahren wie viele Ausreißer identifiziert und wie sehr die Datenvorverarbeitung dadurch für die ML-Aufgabe optimiert wird.

Zusätzlich wird deutlich, dass fehlendes Kontextwissen zum Datensatz die Ausreißeridentifikation deutlich erschwert. Eine wahllose Anwendung von AE-Verfahren auf einen Datensatz ohne Prozesswissen kann zu einer Verschlechterung der Prognose des Modells führen, wie in Kapitel 4.1.2 feststellbar ist. Werden Ausreißer hingegen auf Grundlage von Kontextwissen vor der Anwendung von AE-Verfahren bereits aussortiert, so kann sich eine deutliche Verbesserung der Prognose ergeben, wie in Kapitel 4.1.3 deutlich wird.

Auch die ganzheitliche Betrachtung von Werkstücken, anstelle von einzelnen PDIDs, kann die Ausreißererkennung unterstützen und damit zu einer optimierten Datenvorverarbeitung beitragen. Dies wird anhand des konstruierten Datenraums nachgewiesen und in Kapitel 4.2.2 aufgearbeitet.

Nach der Auswertung der Ergebnisse ergeben sich jedoch auch neue Fragen und Diskussionsgrundlagen. In Kapitel 4.1.3 wird die Vermutung aufgestellt, dass das verwendete Modell imstande ist, Ausreißer zu prognostizieren. Es stellt sich also die Frage, ob ein anderes Modell ggf. weniger robust gegenüber Ausreißern wäre und infolgedessen nicht imstande ist, diese vorherzusagen. Hierdurch würde auch die Anwendung von AE-Verfahren in der Datenvorverarbeitung einen größeren Einfluss auf die Prognose haben. Es könnten möglicherweise also auch eindeutigere Aussagen bezüglich der Eignung und Wirksamkeit der einzelnen Verfahren getroffen werden als in diesem Fall.

Außerdem wurde die These aufgestellt, dass sich mit Prozesswissen vorab bereits Datenpunkte ausschließen lassen, was zu einer Verbesserung der maschinellen Ausreißerentdeckung führt. Dieser Sachverhalt wird in Kapitel 4.1.3 erläutert. Es stellt sich also die Frage, ob mit näherem Hintergrundwissen weitere Muster in den Datensätzen erkannt und dadurch Datenpunkte ausgeschlossen werden können, um die Ausreißererkennung zu optimieren.

5 Zusammenfassung und Ausblick

In diesem Kapitel wird der Bericht zusammengefasst und die Schlussfolgerungen jedes Kapitels logisch zusammengeführt. Anschließend wird ein konkretes Fazit gegeben, welches die relevantesten Erkenntnisse des Berichts umfasst. Abschließend wird ein Ausblick auf weitere Forschungsansätze gegeben.

5.1 Zusammenfassung

Wie in der Einleitung dargelegt, treibt die voranschreitende Digitalisierung die Datenerfassung in verschiedenen Bereichen voran. Dies geschieht aus Eigeninteresse der Unternehmen oder aufgrund gesetzlicher Auflagen. Die wachsende Geschwindigkeit des Datensammelns führt zu einer enormen Datenmenge, die in vielfältigen Unternehmenskontexten genutzt wird. Die steigende Bedeutung der Datenanalyse, insbesondere durch Technologien der Industrie 4.0, betont die Relevanz von Data-Mining-Methoden zur Mustererkennung und -extraktion für die Identifizierung nützlichen Wissens. In diesem Zusammenhang spielt die Datenverarbeitung, im Rahmen von KDD-Modellen, wie des CRISP-DM-Modells, eine zentrale Rolle.

Das CRISP-DM-Modell legt die Grundlage für die iterative Durchführung des Projektes. Der Bericht orientiert sich an diesem Modell, um die ursprüngliche Forschungsfrage zu beantworten. Durch strukturierte Datenvorverarbeitung für die Zeitreihenprognose wird ein Datensatz erstellt, der sich eignet, um die Prognose von Messwerten vorzunehmen. Als Modell wird das Machine-Learning-Modell des XGB verwendet, um zukünftige Werte zu prognostizieren. Eine Prognose zukünftiger Messwerte generiert nützliches Wissen, da diese, wenn sie korrekt prognostiziert sind, Informationen über zukünftige Ausfälle und Werkstückzustände liefern. So ist eine Anwendung für eine Wissensentdeckung im gegebenen Datensatz geschaffen.

In der Evaluation, der Praxisphase werden die Ergebnisse der Zeitreihenprognosen und der analysierten Ausreißererkennungsverfahren dargelegt. Die Evaluation erfolgt mittels Kennzahlen wie MAE und RMSE für die Vorhersagegenauigkeit sowie durch die Betrachtung der Anzahl identifizierter Ausreißer. Die Anwendung von sieben Ausreißererkennungsverfahren und drei Imputationsverfahren auf die ausgewählten Datensätze ermöglicht einen direkten Vergleich der Verfahren. Hierbei wird deutlich, dass die Anzahl der gefundenen Ausreißer von den verwendeten Verfahren abhängig ist, diese jedoch nicht zwangsläufig mit einer verbesserten Prognose einhergehen. Die Resultate zeigen, dass kein eindeutiges Muster hinsichtlich der Ausreißeridentifikation erkennbar ist.

Die Untersuchungen der Ausreißererkennung sowohl auf einzelner Datensatzebene als auch im mehrdimensionalen Datenraum zeigen, dass gezielte Datenvorverarbeitung einen signifikanten Einfluss auf die Prognosegenauigkeit hat. Der Ansatz der ganzheitlichen Ausreißererkennung im Datenraum anhand der Werkstücke verdeutlicht, dass Ausreißer in diesem Kontext nicht immer visuell identifizierbar sind. Dennoch kann gezeigt werden, dass die Anwendung von Ausreißererkennungsverfahren im Datenraum in Kombination mit gezielter Imputation zu einer verbesserten Vorhersage führt. Die präsentierten Ansätze bieten wertvolle Einblicke in die Möglichkeiten der Optimierung von Prognosegenauigkeiten durch gezielte Ausreißererkennung und Imputation. Die Analyse eines möglichen Zusammenhangs zwischen der Vorhersagegenauigkeit (RMSE) und der Anzahl entdeckter Ausreißer zeigt, dass der Pearson-Korrelationskoeffizient für die untersuchten Datensätze lediglich Werte zwischen 0,1 und 0,4 aufweist. Dies deutet darauf hin, dass keine signifikante lineare Korrelation zwischen der Anzahl der identifizierten Ausreißer und dem RMSE besteht. Hierdurch wird deutlich, dass eine erhöhte Anzahl von erkannten Ausreißern nicht zwangsläufig zu einer Verbesserung der Prognose führt.

Die Arbeit zeigt außerdem, dass eine systematische Erkennung und Entfernung von inkonsistenten Werten durch Kontextwissen vor der Modellanwendung die Vorhersagegenauigkeit

steigern kann. Dies wird durch eine Fallunterscheidung mithilfe von Annahmen zu Prozesswissen erreicht, die für den Datensatz PDID-40 entwickelt wurde. Hierbei werden mutmaßliche fehlerhafte Werte vor der Ausreißererkennung entfernt, was zu verbesserten Ergebnissen führt. Dies verdeutlicht die Bedeutung des Einbezugs von Prozesswissen bei der Ausreißerbehandlung und zeigt, dass das „blinde“ Anwenden von Ausreißererkennungsverfahren nicht zielführend ist.

Die Untersuchung der Ausreißererkennung im mehrdimensionalen Datenraum zeigt vielversprechende Ergebnisse. Durch die Identifizierung von ganzen Werkstücken als Ausreißer und die Anwendung von AE-Verfahren im Datenraum wird eine Verbesserung der Prognosegenauigkeit erreicht. Dieser ganzheitliche Ansatz trägt zur effektiven Datenvorverarbeitung für ML-Aufgaben bei. Vergleichend zeigt sich, dass die Einbeziehung von Prozesswissen bei der Ausreißerbehandlung und -erkennung essenziell ist, um bessere Ergebnisse zu erzielen. Diese experimentellen Ansätze bieten somit Möglichkeiten zur Steigerung der Performance von Machine-Learning-Modellen.

Zusammenfassend lässt sich feststellen, dass kein universell optimales Ausreißererkennungsverfahren identifiziert werden konnte. Unter den gezeigten Umständen bestätigen die Ergebnisse dieser Arbeit, die Erkenntnisse von Aggarwal (2017) und Lee et al. (2002), welche in der Einleitung genannt werden. Die Wirksamkeit hängt stark vom jeweiligen Datensatz ab, da verschiedene Verfahren unterschiedlich viele Ausreißer identifizieren und die Datenvorverarbeitung für maschinelles Lernen beeinflussen. Mangelndes Kontextwissen erschwert die Ausreißererkennung, und die gezielte Anwendung von Verfahren auf Grundlage von Kontextwissen verbessert die Prognose. Ein holistischer Ansatz bei der Betrachtung von Werkstücken statt einzelner Merkmale kann die Ausreißererkennung unterstützen. Fragen zur Robustheit von Modellen gegenüber Ausreißern und zur Optimierung der Erkennung durch zusätzliches Wissen bleiben offen.

5.2 Ausblick

Die dargestellten Ergebnisse führen zu weitere Forschungsfragen. Diese lassen sich wie folgt formulieren.

- Eine systematische Untersuchung des Einflusses der Parametrisierung der AE-Verfahren auf die erkannten Ausreißer vornehmen und den resultierenden Einfluss auf die Prognoseergebnisse bestimmen.
- Des Weiteren ist zu prüfen, ob der Ansatz der Fallunterscheidung für inkonsistente Werte auf andere PDIDs übertragbar ist. In diesem Kontext wäre das Einbeziehen von konkreten Informationen über die Prozesse entscheidend, um die Fallunterscheidung verallgemeinern zu können.
- Es ist zu testen, wie sich die weiteren AE-Verfahren auf den Datenraum der Werkstücke und die Prognosen auswirken.
- Im Kontext der Anzahl der gefundenen Ausreißer ist weiter interessant zu untersuchen, wie sich komplexere Imputationsverfahren, wie Glättung oder ML-gestützte Verfahren, auf die Prognoseergbnisse auswirken.

Der vorliegende Bericht zeigt, dass die Identifikation geeigneter Ausreißererkennungsverfahren von der jeweiligen Beschaffenheit der Daten abhängt. Für die gewählte Fragestellung konnte kein universell bestes Verfahren gefunden werden. Zukünftige Untersuchungen könnten sich darauf konzentrieren, datensatzspezifische Merkmale zu identifizieren, die eine Vorhersage darüber ermöglichen, welches Ausreißererkennungsverfahren in einem gegebenen Kontext am effektivsten ist. Weiterhin wäre eine Untersuchung der Auswirkungen verschiedener Kontextinformationen auf die Leistung der Ausreißererkennung von Interesse, um geeignete Strategien für die Nutzung von Prozesswissen zu entwickeln und zu optimieren. Zudem könnte die Untersuchung von hybriden Ansätzen, die verschiedene Ausreißererkennungsverfahren kombinieren, dazu beitragen, robuste Ansätze für die Prognose und Verarbeitung von

Daten mit Ausreißern zu entwickeln. Ein allgemeiner Ansatz würde so einen wertvollen Beitrag für den Erfolg der Durchführung von KDD-Projekten leisten.

6 Literaturverzeichnis

Abedjan, Ziawasch; Golab, Lukasz; Naumann, Felix; Papenbrock, Thorsten (2019): Data profiling. San Rafael, California: Morgan & Claypool Publishers (Synthesis digital library of engineering and computer science, #52). Online verfügbar unter http://ieeexplore.ieee.org/servlet/opac?bknumber=8540360.

Aggarwal, Charu C. (Hg.) (2015): Data Mining. Cham: Springer International Publishing.

Aggarwal, Charu C. (2017): Outlier analysis. Second edition. Cham: Springer.

Alexandru Prisacaru; Ernesto Oquelis Guerrero; Przemyslaw Jakub Gromala; Bongtae Han; Guo Qi Zhang (2019): 2019 20th International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE). Piscataway, NJ: IEEE. Online verfügbar unter https://ieeexplore.ieee.org/servlet/opac?punumber=8718745.

Aljaž, Ferencek; Mirjana Kljajić, Borštnar (2020): Data quality assessment in product failure prediction models, S. 79–86.

Angelov, P.; Atanassov, K. T.; Doukovska, L.; Hadjiski, M.; Jotsov, V.; Kacprzyk, J. et al. (Hg.) (2015): Intelligent Systems'2014. Proceedings of the 7th IEEE International Conference Intelligent Systems IS'2014, September 24-26, 2014, Warsaw, Poland, Volume 1 Mathematical Foundations, Theory, Analyses. 1st ed. 2015. Cham: Springer International Publishing; Imprint: Springer (Advances in Intelligent Systems and Computing, 322).

Bamberg, Günter; Baur, Franz; Krapp, Michael (2017): Statistik. Eine Einführung für Wirtschafts- und Sozialwissenschaftler. 18., vollständig aktualisierte Auflage. Berlin/Boston: De Gruyter Oldenbourg.

Batini, Carlo; Cappiello, Cinzia; Francalanci, Chiara; Maurino, Andrea (2009): Methodologies for data quality assessment and improvement. In: *ACM Comput. Surv.* 41 (3), S. 1–52. DOI: 10.1145/1541880.1541883.

Batini, Carlo; Scannapieca, Monica (2006): Data Quality. Concepts, Methodologies and Techniques. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg (Data-Centric Systems and Applications). Online verfügbar unter https://link.springer.com/book/10.1007/3-540-33173-5.

Bertolini, Massimo; Mezzogori, Davide; Neroni, Mattia; Zammori, Francesco (2021): Machine Learning for industrial applications: A comprehensive literature review. In: *Expert Systems with Applications* 175, S. 114820. DOI: 10.1016/j.eswa.2021.114820.

Bhattacharya, Gautam; Ghosh, Koushik; Chowdhury, Ananda S. (2015): Outlier detection using neighborhood rank difference. In: *Pattern Recognition Letters* 60-61, S. 24–31. DOI: 10.1016/j.patrec.2015.04.004.

Bilal, Mehwish; Ali, Ghulam; Iqbal, Muhammad Waseem; Anwar, Muhammad; Malik, Muhammad Sheraz Arshad; Kadir, Rabiah Abdul (2022): Auto-Prep: Efficient and Automated Data Preprocessing Pipeline. In: *IEEE Access* 10, S. 107764–107784. DOI: 10.1109/ACCESS.2022.3198662.

Blake, Roger; Mangiameli, Paul (2011): The Effects and Interactions of Data Quality and Problem Complexity on Classification. In: *J. Data and Information Quality* 2 (2), S. 1–28. DOI: 10.1145/1891879.1891881.

Booysen, W.; Botes, L. A.; Hamer, W. (2017): A practical methodology for the systematic identification of outliers. In: 2017 International Conference on the Industrial and Commercial Use of Energy (ICUE). 2017 International Conference on the Industrial and Commercial Use of Energy (ICUE). Cape Town, South Africa, 15.08.2017 - 16.08.2017: IEEE, S. 1–6.

Breunig, M. M.; Kriegel, H.; Ng, R. T.; Sander, J. (2000): LOF: Identifying Density-Based Local Outliers. In: *Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX*, S. 93–104.

Carmona, Jesus; Lopez, Ivan; Mateo, Josep; Jimenez, Laureano; Aldana, Edwyn (2020): A distance-based method for outlier detection on high dimensional datasets. In: *IEEE Latin Am. Trans.* 18 (03), S. 589–597. DOI: 10.1109/TLA.2020.9082731.

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

Chapmann, Pete; Clinton, Julian; Kerber, Randy; Khabaza, Thomas; Reinartz, Thomas; Shearer, Colin; Wirth, Rüdiger (2000): CRISP-DM 1.0. Step-by-step data mining guide. SPSS. Online verfügbar unter https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72.

Chen, Jian; Wu, Yefan; Lin, Zhiyong; Zhao, Liang; Wang, Qing; Hou, Hui; Deng, Xiangtian (2021): Review of Load Forecasting Based on Artificial Intelligence Models. In: 2021 6th Asia Conference on Power and Electrical Engineering (ACPEE). 2021 6th Asia Conference on Power and Electrical Engineering (ACPEE). Chongqing, China, 08.04.2021 - 11.04.2021: IEEE, S. 340–344.

Chheda, Vatsal; Kapadia, Samit; Lakhani, Bhavya; Kanani, Pratik (2021): Automated Data Driven Preprocessing and Training of Classification Models. In: 2021 4th International Conference on Computing and Communications Technologies (ICCCT). 2021 4th International Conference on Computing and Communications Technologies (ICCCT). Chennai, India, 16.12.2021 - 17.12.2021: IEEE, S. 27–32.

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

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

Dani, Yasi; Gunawan, Agus Yodi; Indratno, Sapto Wahyu (2022): Detecting Online Outlier for Data Streams using Recursive Residual. In: 2022 Seventh International Conference on Informatics and Computing (ICIC). 2022 Seventh International Conference on Informatics and Computing (ICIC). Denpasar, Bali, Indonesia, 08.12.2022 - 09.12.2022: IEEE, S. 1–7.

Desai, Vinod; Dinesha, H. A. (2020): A Hybrid Approach to Data Pre-processing Methods. In: 2020 IEEE International Conference for Innovation in Technology (INOCON). 2020 IEEE International Conference for Innovation in Technology (INOCON). Bangluru, India, 06.11.2020 - 08.11.2020: IEEE, S. 1–4.

Diniz, H.; Andrade, L. de; Carvalho, A. de; Andrade, M. de (1999): Architecture design of artificial neural networks based on Box & Jenkins models for time series prediction. In: Proceedings Third International Conference on Computational Intelligence and Multimedia Applications. ICCIMA'99 (Cat. No.PR00300). Third International Conference on Computational Intelligence and Multimedia Applications. ICCIMA'99. New Delhi, India, 23-26 Sept. 1999: IEEE Comput. Soc, S. 29–34.

Diop, Mouhamed; Mamadou Samba CAMARA; Ibrahima FALL; Alassane BAH (2017): A Methodology for Prior Management of Temporal Data Quality in a Data Mining Process. 2017 Intelligent Systems and Computer Vision (ISCV) : April 17-19, 2017, Faculty of Sciences Dhar El Mahraz (FDSM), Fez, Morocco. Online verfügbar unter http://ieeexplore.ieee.org/servlet/opac?punumber=8048780.

Domanski, Pawel D. (2020): Statistical outlier labelling – a comparative study. In: 2020 7th International Conference on Control, Decision and Information Technologies (CoDIT). 2020

7th International Conference on Control, Decision and Information Technologies (CoDIT). Prague, Czech Republic, 29.06.2020 - 02.07.2020: IEEE, S. 439–444.

Dost, Shahi; Anwer, Sajid; Saud, Faryal; Shabbir, Maham (2017): Outliers classification for mining evolutionary community using Support Vector Machine and Logistic Regression on Azure ML. In: 2017 International Conference on Communication, Computing and Digital Systems (C-CODE). 2017 International Conference on Communication, Computing and Digital Systems (C-CODE). Islamabad, Pakistan, 08.03.2017 - 09.03.2017: IEEE, S. 216–221.

Ester; Martin; Kriegel; Hans-Peter; Sander; Jorg et al.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: *Proc. of the 2nd ACM Intl. Conf on Knowledge Discovery and Data Mining (KDD)*, 226-231.

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

Ferdowsi, H.; Jagannathan, S.; Zawodniok, M. (2013): A neural network based outlier identification and removal scheme. In: 2013 IEEE Conference on Prognostics and Health Management (PHM). 2013 IEEE Conference on Prognostics and Health Management (PHM). Gaithersburg, MD, USA, 24.06.2013 - 27.06.2013: IEEE, S. 1–6.

Ferdowsi, Hasan; Jagannathan, Sarangapani; Zawodniok, Maciej (2014): An online outlier identification and removal scheme for improving fault detection performance. In: *IEEE transactions on neural networks and learning systems* 25 (5), S. 908–919. DOI: 10.1109/TNNLS.2013.2283456.

Freitag, M.; Kück, M.; Alla, A. A.; Lütjen, M. (2015): Potenziale von Data Science in Produktion und Logistik: Teil 2 - Vorgehensweise zur Datenanalyse und Anwendungsbeispiele. In: *Industrie 4.0 Management* (35), S. 39–46.

Frigui, H. (2004): Pre-processing for data clustering. In: IEEE Annual Meeting of the Fuzzy Information, 2004. Processing NAFIPS '04. IEEE Annual Meeting of the Fuzzy Information, 2004. Processing NAFIPS '04. Banff, Alta., Canada, 27.06.2004 - 30.06.2004: IEEE, 967-972 Vol.2.

García, Salvador; Luengo, Julián; Herrera, Francisco (2015): Data preprocessing in data mining. Cham: Springer (Intelligent Systems Reference Library, 72). Online verfügbar unter http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=839026.

Garg, Nikhil; Singh, Sandeep Kumar (2021): Machine Learning based Forecasting of Wind Power. In: 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT). 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT). Bhopal, India, 18.06.2021 - 19.06.2021: IEEE, S. 612–616.

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

Han, Jianchao; Rodriguez, Juan C.; Beheshti, Mohsen (2008): Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner. In: 2008 Second International Conference on Future Generation Communication and Networking. 2008 Second International Conference on Future Generation Communication and Networking (FGCN). Hainan, China, 13.12.2008 - 15.12.2008: IEEE, S. 96–99.

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

Imtiaz Ahmed; Aldo Dagnino; Alessandro Bongiovi; Yu Ding (2018): 2018 IEEE 14th International Conference on Automation Science and Engineering (CASE). 20-24 Aug. 2018. Online verfügbar unter http://ieeexplore.ieee.org/servlet/opac?punumber=8536777.

Jayaramulu, C.; Venkateswarlu, Bondu (2022): DLOT-Net: A Deep Learning Tool For Outlier Identification. In: 2022 6th International Conference on Electronics, Communication and Aerospace Technology. 2022 6th International Conference on Electronics, Communication and Aerospace Technology (ICECA). Coimbatore, India, 01.12.2022 - 03.12.2022: IEEE, S. 1143–1147.

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

Laaroussi, Houria; Guerouate, Fatima; sbihi, Mohamed (2020): Deep Learning Framework for Forecasting Tourism Demand. In: IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD). IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD). Marrakech, Morocco, 24.11.2020 - 27.11.2020: IEEE, S. 1–4.

Larose, Daniel T.; Larose, Chantal D. (2014a): Discovering knowledge in data. An introduction to data mining. 2. ed. Hoboken, NJ: Wiley (Wiley series on methods and applications in data mining).

Larose, Daniel T.; Larose, Chantal D. (Hg.) (2014b): Discovering Knowledge in Data. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Lee, Y. W.; Strong, D. M.; Kahn, B. K.; Wang, R. Y. (2002): AIMQ: a methodology for information quality assessment. Issue 2. Volume 40: Information & Management. Online verfügbar unter https://doi.org/10.1016/S0378-7206(02)00043-5.

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

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

Mandhare, H. C.; Idate, Prof. S. R. Idate (2017): A Comparative Study of Cluster Based Outlier Detection, Distance Based Outlier Detection and Density Based Outlier Detection Techniques. Piscataway, NJ: IEEE. Online verfügbar unter http://ieeexplore.ieee.org/servlet/opac?punumber=8241057.

Mariscal, Gonzalo; Marbán, Óscar; Fernández, Covadonga (2010): A survey of data mining and knowledge discovery process models and methodologies. In: *The Knowledge Engineering Review* 25 (2), S. 137–166. DOI: 10.1017/S0269888910000032.

Martinez-Plumed, Fernando; Contreras-Ochando, Lidia; Ferri, Cesar; Hernandez Orallo, Jose; Kull, Meelis; Lachiche, Nicolas et al. (2020): CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. In: *IEEE Trans. Knowl. Data Eng.* (33), Artikel 8, S. 1–15. DOI: 10.1109/TKDE.2019.2962680.

Masood, Sheikh Wakie; Begum, Shahin Ara (2022): Comparison of Resampling Techniques for Imbalanced Datasets in Student Dropout Prediction. In: 2022 IEEE Silchar Subsection Conference (SILCON). 2022 IEEE Silchar Subsection Conference (SILCON). Silchar, India, 04.11.2022 - 06.11.2022: IEEE, S. 1–7.

Mertens, Peter; Rässler, Susanne (2012): Prognoserechnung. Heidelberg: Physica-Verlag HD.

Moreno-Sanchez, Pedro A. (2020): Features Importance to Improve Interpretability of Chronic Kidney Disease Early Diagnosis. In: 2020 IEEE International Conference on Big Data (Big Data). 2020 IEEE International Conference on Big Data (Big Data). Atlanta, GA, USA, 10.12.2020 - 13.12.2020: IEEE, S. 3786–3792.

Moroff, Nikolas Ulrich; Kurt, Ersin; Kamphues, Josef (2021): Machine Learning and Statistics: A Study for assessing innovative Demand Forecasting Models. In: *Procedia Computer Science* 180, S. 40–49. DOI: 10.1016/j.procs.2021.01.127.

Muthukrishnan, S.; Shah, R.; Vitter, J. S. (2004): Mining deviants in time series data streams. In: Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004. Santorini Island, Greece, 21-23 June 2004: IEEE, S. 41–50.

Nematzadeh, Zahra; Ibrahim, Roliana; Selamat, Ali (2020): A hybrid model for class noise detection using k-means and classification filtering algorithms. In: *SN Appl. Sci.* 2 (7). DOI: 10.1007/s42452-020-3129-x.

Nickerson, Robert; Muntermann, Jan; Varshney, Upkar; Isaac, Henri (2009): Taxonomy Development In Information Systems: Developing A Taxonomy Of Mobile Applications. In: *HAL, Working Papers*.

Nino, Mikel; Blanco, Jose Miguel; Illarramendi, Arantza (2015): Business understanding, challenges and issues of Big Data Analytics for the servitization of a capital equipment manufacturer. In: 2015 IEEE International Conference on Big Data (Big Data). 2015 IEEE International Conference on Big Data (Big Data). Santa Clara, CA, USA, 29.10.2015 - 01.11.2015: IEEE, S. 1368–1377.

Osborne, Jason W.; Overbay, Amy (2004): The power of outliers (and why researchers should ALWAYS check for them).

Pearson, R. K. (2002): Outliers in process modeling and identification. In: *IEEE Transactions on Control Systems Technology* 2002 (10), S. 55–63.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O. et al. (2011): Scikit-learn: Machine Learning in Python. In: *Journal of Machine Learning Research* 12, S. 2825–2830. Online verfügbar unter https://scikit-learn.org/stable/about.html#citing-scikit-learn, zuletzt geprüft am 18.08.2023.

Plaue, Matthias (2021): Data Science. Grundlagen, Statistik und maschinelles Lernen. Berlin, Heidelberg: Springer Spektrum (Lehrbuch).

Raheem, Nasir (2019): Big data. A tutorial-based approach. Boca Raton, London, New York: CRC Press Taylor & Francis Group (ProQuest Ebook Central). Online verfügbar unter http://site.ebrary.com/lib/alltitles/docDetail.action?docID=11656330.

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

Ray, Susmita (2019): A Quick Review of Machine Learning Algorithms. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). Faridabad, India, 14.02.2019 - 16.02.2019: IEEE, S. 35–39.

Runkler, Thomas A. (2016): Data analytics. Models and algorithms for intelligent data analysis. 2nd edition. Wiesbaden: Springer Vieweg (Lehrbuch). Online verfügbar unter http://www.springer.com/.

Saltz, Jeffrey S. (2021): CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. In: 2021 IEEE International Conference on Big Data (Big Data). 2021 IEEE International Conference on Big Data (Big Data). Orlando, FL, USA, 15.12.2021 - 18.12.2021: IEEE, S. 2337–2344.

Schröer, Christoph; Kruse, Felix; Gómez, Jorge Marx (2021): A Systematic Literature Review on Applying CRISP-DM Process Model. In: *Procedia Computer Science* 181, S. 526–534. DOI: 10.1016/j.procs.2021.01.199.

Sharma, Sumana; Osei-Bryson, Kweku-Muata (2008): Organization-Ontology Based Framework for Implementing the Business Understanding Phase of Data Mining Projects. In: Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008). 2008 41st Annual Hawaii International Conference on System Sciences. Waikoloa, HI, 07.01.2008 - 10.01.2008: IEEE, S. 77.

Smith, Michael R.; Martinez, Tony (2011): Improving classification accuracy by identifying and removing instances that should be misclassified. In: The 2011 International Joint Conference on Neural Networks. 2011 International Joint Conference on Neural Networks (IJCNN 2011 - San Jose). San Jose, CA, USA, 31.07.2011 - 05.08.2011: IEEE, S. 2690–2697.

Statista (2023). Online verfügbar unter https://de.statista.com/statistik/daten/studie/257988/umfrage/prognose-zum-umsatz-mit-big-data-loesungen-weltweit-nach-segment, zuletzt geprüft am 27.08.2023.

Sumana, Sharma; Kweku-Muata, Osei-Bryson (2010): Toward an integrated knowledge discovery and data mining process model.

vom Brocke, Jan; Simons, Alexander; Niehaves, Björn; Riemer, Kai; Plattfaut, Ralf; Cleven, Anne (2009): Reconstructing the giant: On the importance of rigour in documenting the literature search process. In: European Conference on Information Systems.

vom Brocke, Jan; Simons, Alexander; Riemer, Kai; Niehaves, Björn; Plattfaut, Ralf; Cleven, Anne (2015): Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. In: *CAIS* 37. DOI: 10.17705/1CAIS.03709.

Wan, Fangyi; Guo, Gaodeng; Zhang, Chunlin; Guo, Qing; Liu, Jie (2019): Outlier Detection for Monitoring Data Using Stacked Autoencoder. In: *IEEE Access* 7, S. 173827–173837. DOI: 10.1109/ACCESS.2019.2956494.

Wang, Richard Y.; Strong, Diane M. (1996): Beyond Accuracy: What Data Quality Means to Data Consumers. In: *Journal of Management Information Systems* 12 (4), S. 5–33. DOI: 10.1080/07421222.1996.11518099.

Webster, Jane; Watson, Richard T. (2002): Analyzing the Past to Prepare for the Future: Writing a Literature Review (26), Artikel 2, S. 13–23. Online verfügbar unter http://www.jstor.org/stable/4132319., zuletzt geprüft am 09.08.2023.

Widiputra, Harya; Mailangkay, Adele; Gautama, Elliana (2020): Time-Series Outliers Detection Algorithm with Clustering Approach on Non-Linear Trends. In: 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE). 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE). Yogyakarta, Indonesia, 15.09.2020 - 16.09.2020: IEEE, S. 25–30.

Wirth, Rüdiger; Hipp, Jochen (2000a): CRISP-DM Towards a Standard Process Model for Data Mining. In: Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, S. 29–39.

Wirth, Rüdiger; Hipp, Jochen (2000b): CRISP-DM Towards a Standard Process Model for Data Mining. In: Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, S. 29–39.

Xu, Chaofan; Tang, Jian; Sun, Zijian; Xia, Heng; Xu, Zhe; Xu, Wen (2021): Multi-window Drift Detection Method Based on Integrating Outlier Identification and Input/Output Space Information with Its Application. In: 2021 China Automation Congress (CAC). 2021 China Automation Congress (CAC). Beijing, China, 22.10.2021 - 24.10.2021: IEEE, S. 7206–7211.