technische universität dortmund

itpl IT in Produktion und Logistik

Technische Universität Dortmund

Fakultät für Maschinenbau

Fachgebiet für IT in Produktion und Logistik (ITPL)

Masterarbeit

von

Alexander Schmidt, B. Sc.

Studiengang: Logistik

Matr. Nr.: 181772

Optimierung des Entscheidungsprozesses zur Einlagerung von Produkten in einem teilautomatisierten Logistikzentrum unter Anwendung von Data Mining

ausgegeben am:
30.05.2016

eingereicht am:
14.11.2016

Erstprüfer:	Prof. Dr.-Ing. Markus Rabe	(TU Dortmund, ITPL)
Zweitprüfer:	Dr. Jens Derner	(SSI Schäfer Noell GmbH)

Kurzfassung

In dieser Arbeit wird der Einlagerungsprozess eines teilautomatisierten Logistikzentrums optimiert. Für die erfolgreiche Optimierung werden zur Einführung die Grundlagen zur Einlagerung erläutert. Zur Lösung der definierten Problemstellung werden Data-Mining-Verfahren benötigt, welche im Grundlagenteil der Arbeit vorgestellt werden. Neben den Data-Mining-Verfahren werden Möglichkeiten zur Vorverarbeitung der Daten näher vorgestellt. Zusätzlich werden Vergleichsmöglichkeiten der Data-Mining-Verfahren aufgezeigt.

Zur Optimierung des Einlagerungsprozesses werden zwei Handlungsalternativen entwickelt, welche beide als ein Vorgehensmodell zur Anwendung von Data Mining zu verstehen sind. Mit der ersten Handlungsalternative wird der Lagerbereich für das jeweilige Produkt bestimmt. Bei der Anwendung der zweiten Handlungsalternative wird die Wahrscheinlichkeit für die Auslagerung der Produkte vorausgesagt. Der ermittelte Wert kann in verschiedenen Algorithmen genutzt werden, welche den Lagerbereich bestimmen. Beide Handlungsalternativen werden anhand unterschiedlicher Einflussfaktoren verglichen. Dabei wird eine Entscheidung zugunsten der zweiten Handlungsalternative getroffen.

Die zweite Handlungsalternative wird mit Hilfe des Programmes Rapidminer prototypisch umgesetzt. Dabei werden Methoden zur Reduktion der Datenmengen vorgestellt. Auf Basis der reduzierten Daten wird ein Neuronales Netz entwickelt, welches sich auf zukünftige Daten anwenden lässt. Im Ergebnis dieser Arbeit kann festgehalten werden, dass Data-Mining genutzt werden kann, um ein Einlagerungsproblem zu optimieren.

Abstract

In this thesis the storage process of a partially automated logistics center will be optimized. For a successful optimization, the basics of storage will be defined during the first part of the literature research. To solve the defined problem, data mining methods are required, which are presented in the research part of the thesis. In addition to the data mining methods, possibilities for the data preprocessing are explained particularly and comparative possibilities of the data mining methods are presented.

Two different alternatives are developed to optimize the storage process, both are understood as a model of Knowledge in Discovery Databases. The storage area for the respective product is determined with the first alternative. In the application of the second alternative, the probability of the products being outsourced is determined. This can be used in various algorithms which determine the storage area. Both alternatives are compared by different influencing factors. A decision is made to the credit of the second alternative.

The second alternative is implemented prototypically by using Rapidminer. Afterwards different techniques of reducing the amount of data are presented. On the reduced data, a neural network is developed, which can be applied to future data. As a result of this thesis it can be stated that data mining can be used to optimize a storage problem.

Sperrvermerk

Die vorgelegte Masterarbeit basiert auf internen, vertraulichen Daten und Informationen des Unternehmens SSI Schäfer Noell. Der Anhang dieser Abschlussarbeit darf nur Personen zugänglich gemacht werden, welche die verpflichtende Geheimhaltungserklärung unterzeichnet haben. Eine Veröffentlichung und Vervielfältigung des Anhangs ist - auch in Auszügen - nicht gestattet. Eine Einsichtnahme des Anhangs durch Unbefugte bedarf einer ausdrücklichen Genehmigung durch den Verfasser und das Unternehmen.

Inhaltsverzeichnis

1 Einleitung

Der Onlinehandel gewinnt heutzutage permanent an Bedeutung und ist im Alltag nicht mehr wegzudenken. Immer mehr Anbieter versuchen sich auf dem Markt erfolgreich zu etablieren. Die Wettbewerbsfähigkeit der Anbieter ist unter anderem abhängig von der Effektivität ihrer Logistikzentren. Das Ziel dieser Logistikzentren ist, zu möglichst geringen Kosten die Produkte zu lagern, zu kommissionieren (engl. = to pick) und zu verpacken.

Durchschnittlich kostet die Logistikleistung in einem Logistikzentrum zwischen drei und zwölf Euro pro Bestellung, abhängig von der Anzahl der Produkte [VKS12]. Dabei ist die Kommissionierarbeit der Kostentreiber der Logistikleistung mit einer Spanne zwischen 0,75 und 6,25 Euro abhängig von der Anzahl der Bewegungen des Produktes [VKS12]. Das Ziel ist, diese Kosten möglichst gering zu halten. Die Reduzierung der Bewegungen der Produkte lässt sich mit Hilfe von automatisierten Lagern realisieren, wodurch die Kosten für die Kommissionierarbeit gesenkt werden können. Ein automatisiertes Lager bietet eine geringere Flexibilität, als ein manuelles Lager. Aufgrund der derzeitigen Technik und der großen Produktvielfalt ist es nicht möglich ein automatisiertes Lager im Onlinehandel umzusetzen. Deswegen wird eine Kombination aus einem manuellen Lager und automatisierten Lager benötigt. Ein teilautomatisiertes Lager verbindet die Vor-und Nachteile der beiden anderen Lagertypen in einem System und eignet sich für den Einsatz im Onlinehandel.

In Zusammenarbeit zwischen einem Onlinehändler und der SSI Schäfer Noell GmbH wurde ein solches teilautomatisiertes Logistikzentrum umgesetzt. Die benötigten Produkte werden aus zwei verschiedenen Lagern zum Kommissionierplatz gefördert. Dabei stellt eine Fachbodenregalanlage den manuellen Teil und ein automatisches Hochregallager den automatisierten Teil des Logistikzentrums dar. Das manuelle Picksystem ist sehr arbeitskraftabhängig. Je mehr Menschen dort arbeiten, desto höher ist die Ausbringungsmenge für das gesamte Logistikzentrum. Diese Anpassung an die Auslastung ist wichtig im Onlinehandel, weil dieser über das Jahr verteilt starken Schwankungen unterliegt. Deswegen soll in den Niedriglastzeiten das manuelle Lager auf ein Minimum reduziert werden, um die Logistikkosten für die Kommissionierarbeit gering zu halten. Im Gegensatz dazu soll in der Niedriglastzeit aus dem automatisierten Lager ein wesentlich höherer Anteil zu den Ware-zu-Person Arbeitsplätzen gefördert werden. Die Produkte werden aus beiden Lagern in wiederverwertbaren Kisten zu Ware-zu-Person-Arbeitsplätzen gefördert und dort mit Hilfe eines Sorters in die verschiedenen Bereiche zur Auftragskonsolidierung sortiert.

Die Entscheidung in welchen der beiden Bereiche das Produkt eingelagert werden soll, ist im Moment nicht optimal. Die Produkte werden hauptsächlich im manuellen Lager eingelagert, wobei die Leistungsgrenze des automatisierten Lagers nicht ausgeschöpft wird. Deshalb stellt der Entscheidungsprozess, wo das jeweilige Produkt eingelagert werden soll, das Kernproblem dieser Arbeit dar.

Das Ziel dieser Arbeit ist, die Entscheidungsfindung für die Einlagerung zu optimieren. Dafür sollen Vergangenheitsdaten mit Data-Mining-Verfahren analysiert werden. Eine ganzheitliche Betrachtung des Problems erfordert es, ein bereits definiertes Vorgehensmodell zur Durchführung

von Data-Mining-Verfahren anzuwenden. Es gilt einen allgemeingültigen Ablauf, unter zu Hilfenahme eines definierten Vorgehensmodelles, zur Optimierung des Einlagerungsprozesses zu finden. Insbesondere müssen aufgrund der großen Anzahl an unterschiedlichen Produkten konzeptionelle Lösungen zur Beherrschbarkeit der Produktvielfalt in den Vorverarbeitungsphasen des Data Minings (DM) realisiert werden. Um die Allgemeingültigkeit nachzuweisen, wird eine prototypische Anwendung auf Vergangenheitsdaten durchgeführt. Die Ergebnisse des Modells werden als Grundlage für das Slotting genutzt. Das Slotting stellt eine intelligente Lagerplatzverwaltung dar, welche sich auf Zukunftsprognosen stützt [RCB14]. Die Umsetzung und Implementierung in die bestehende Software ist nicht Bestandteil dieser Arbeit.

Um zu einer erfolgreichen prototypischen Umsetzung des beschriebenen Vorgehensmodells zu gelangen, muss ein Vorgehensmodell zur Anwendung der DM-Verfahren definiert werden. In Kapitel 3 wird das Vorgehensmodell nach [FPS96] beschrieben und vorgestellt. Das Modell enthält neun verschiedene Schritte. In dieser Arbeit sind drei dieser Schritte besonders hervorzuheben, bearbeitet werden jedoch alle neun Schritte. Die Vorverarbeitungsmethoden werden als ein Teilschritt des Vorgehensmodells in Kapitel 3 erläutert. Die Auswahl von geeigneten Vorverarbeitungsmethoden benötigt ein Wissen über die in Kapitel 2 beschriebenen Herausforderungen eines teilautomatisierten Logistikzentrums. Im Vorgehensmodell folgt nach der Datenvorverarbeitung das DM. Ein weiterer Teilschritt ist das DM, dies stellt eine Vielzahl von Werkzeugen zur Verfügung, welche auf die korrekt vorverarbeiteten Daten angewendet werden können. Eine Eingrenzung der Werkzeuge kann durch die in Kapitel 2 beschriebenen Grundlagen zu Entscheidungen getroffen werden. Auf Basis dieser Eingrenzung werden drei verschiedene Verfahren zur Klassifikation beschrieben. Die Untersuchung der Anwendung auf die Problemstellung wird in Kapitel 4 vorgenommen, hierbei werden die Ergebnisse von drei verschiedenen Verfahren verglichen (Handlungsalternative Eins). Zu diesem Vergleich werden geeignete Vergleichsverfahren benötigt, welche am Ende von Kapitel 3 beschrieben werden. Die Vergleichsverfahren stellen den letzten Teilschritt des Vorgehensmodells dar. Eine weitere Möglichkeit (Handlungsalternative Zwei) zur Anwendung von DM zur Lagerplatzvergabe ergibt sich aus dem Abschnitt zur Einlagerung. Hierbei werden verschiedene Einlagerungsstrategien erläutert und Strategien, bei denen die Prognose der Häufigkeit zur Anwendung kommt. Neben den Einlagerungsstrategien werden die Prozesse in einem Logistikzentrum erläutert, um ein besseres Verständnis für das Problem zu ermöglichen. Das DM bietet Werkzeuge zur Prognose der Häufigkeit. Eines dieser Werkzeuge wird ebenfalls in Kapitel 3 beschrieben. Der ermittelte Prognosewert muss daraufhin in einen vorher feststehenden Algorithmus integriert werden um den genauen Lagerbereich vorhersagen zu können. Dieser Algorithmus wird in Zusammenhang mit der Entwicklung des Vorgehensmodells in Kapitel 4 beschrieben. Somit entsteht ein weiterer Ansatz zur Lösung des vorliegenden Entscheidungsproblems. Diese beiden Handlungsalternativen der Lösung werden nun in Kapitel 4 in ein allgemeingültiges Modell überführt. Nach der erfolgreichen Überführung werden beide Handlungsalternativen ausführlich erläutert und zum Schluss jeweils auf ihre Implementierungsmöglichkeit und Anwendungsmöglichkeit kontrolliert. Nun wird sich aus Handlungsalternative Eins für ein DM-Verfahren entschieden und die Ergebnisse der jeweiligen Verfahren mit der zweiten Handlungsalternative verglichen. Anschließend wird eine Entscheidung für eine der beiden Handlungsalternativen getroffen. Nach der Entscheidung für eine der beiden Handlungsalternativen wird für die ausgewählte Alternative eine prototypische Umsetzung durchgeführt. Das

Ergebnis stellt entweder den prognostizierten Lagerbereich dar (Handlungsalternative Eins) oder einen Wert zur Prognose, ob das Produkt in der kommenden Woche ausgelagert wird (Handlungsalternative Zwei). Eine Zusammenfassung der erarbeiteten Ergebnisse wird im letzten Kapitel präsentiert. Zuletzt beinhaltet diese Arbeit einen Ausblick auf die kommende Implementierung und Möglichkeiten zur Validierung des entwickelten Modells.

2 Grundlagen zur Einlagerung

In diesem Kapitel werden die logistischen Grundlagen für ein teilautomatisiertes Logistikzentrum gelegt, die Herausforderungen der Produktvielfalt erläutert und der Entscheidungsprozess definiert. Im ersten Abschnitt werden die verschiedenen Bereiche in einem Lagersystem vorgestellt. Hierbei wird in Bezug auf die Problemstellung die Einlagerung näher betrachtet. Detailliert wird auf vorhandene Einlagerungsstrategien eingegangen. Anschließend werden der Entscheidungsprozess und dessen Anwendung von DM näher erläutert. Der letzte Abschnitt liefert Informationen über die Herausforderungen der Produktvielfalt im Kontext des Onlinehandels.

2.1 Prozessschritte eines Logistikzentrums

Die Bereiche in einem Lagersystem lassen sich in verschiedene Prozesse gliedern. Insgesamt kann zwischen sieben verschiedenen Bereichen unterschieden werden, angefangen bei der Warenvereinnahmung bis hin zum Versand des Produktes. Jedes teilautomatisierte Logistikzentrum besitzt einen individuellen Aufbau und Ablauf. Folgende Bereiche stellen einen Standardablauf dar. Dieser Standardablauf ist bei großen Systemen unabdingbar, da dieser sich nahtlos in die jeweiligen Supply Chains einfügen soll [HS08]. Im Folgenden werden die Bereiche gesammelt vorgestellt, wobei der Bereich der Einlagerung in einem gesonderten Abschnitt beschrieben wird.

2.1.1 Standardisierte Bereiche in einem Logistikzentrum

In einem Logistikzentrum existieren verschiedene Bereiche mit unterschiedlichen Aufgaben. Folgende Bereiche können allgemeingültig für alle Lagerarten definiert werden [Mar14]:

- Wareneingangssystem
- Zuführendes Transportsystem
- Einheiten- und/oder Kommissionierlager mit
    - Einlagerungssystem
    - Lagerungssystem als Boden und/oder Regallagerung
    - Auslagerungssystem
- Abführendes Transportsystem
- Warenausgangssystem

Im Bereich des *Wareneingangssystems* werden im ersten Schritt die Produkte vom Lieferanten angenommen. Nachdem die Produkte entladen sind, werden Menge und Qualität kontrolliert. Sofern das entladene Produkt das erste Mal im Logistikzentrum eingelagert wird, müssen alle relevanten Stammdaten des Produktes aufgenommen werden. Zu den relevanten Daten zählen unter anderem das Gewicht, die Abmessungen und die Anzahl der Produkte in einer Verpackungseinheit. Weiterhin muss entschieden werden, ob die Produkte zu Lagereinheiten zusammengefasst werden müssen oder ob sie in ihrer ursprünglichen Verpackungseinheit eingelagert werden können [HS08].

Das zuführende und abführende *Transportsystem* kann manuell, teil- oder vollautomatisch aufgebaut sein. Als Hauptaufgabe ist eine Verbindung zwischen dem Wareneingangssystem und

dem Lager herzustellen. Sofern das Produkt benötigt wird, stellt das Transportsystem die Verbindung zwischen Lager und Warenausgangsystem dar.

Im nächsten Bereich dem *Einheiten- und/oder Kommissionierlager* sind drei Hauptaufgaben zu erfüllen. Bei der ersten Aufgabe muss das Produkt eingelagert werden. Dieses wird detaillierter im folgenden Abschnitt beschrieben. Die zweite Aufgabe ist das Lagern des Produktes. Die Auslagerung der Produkte ist die letzte Aufgabe. Dabei erfolgt die Auslagerung auf Anweisung des Lagerverwaltungssystems. Die Auslagerung der Produkte kann nach unterschiedlichen Strategien erfolgen. Zwei der bekanntesten lauten: First-In-First-Out (FIFO) und Last-In-First-Out (LIFO). Eine Vielzahl an weiteren Auslagerungsstrategien existiert noch. Sie werden an dieser Stelle jedoch nicht näher erläutert [HS08].

Der letzte Bereich ist das *Warenausgangssystem*. Die ankommenden Güter müssen je nach Erforderlichkeit noch konsolidiert werden. Zusammengehörende Produkte müssen zum jeweiligen Auftrag zusammengefasst werden, um versandbereit zu sein. Sind die Produkte zusammengefasst, werden sie verpackt, beschriftet und zum Versandbereich übergeben. Dort werden sie abhängig vom gewählten Transportmittel sortiert und an den Dienstleister für den Versand übergeben [Mar14].

2.1.2 Einlagerung

Die Einlagerung stellt das Kernproblem in dieser Arbeit dar und wird detaillierter betrachtet. Die Einlagerung kann in zwei unterschiedliche Bereiche unterteilt werden. Das ist zum einen die Bestimmung des Lagerplatzbereiches und zum anderen die Strategien um den Lagerplatz in dem vorher ausgewählten Lagerbereich zu vergeben. Es existieren verschiedene Möglichkeiten den Lagerplatz oder den Lagerplatzbereich zu bestimmen. Neben der allgemeinen Erläuterung werden verschiedene Strategien zur Lagerplatzvergabe aus der Literatur betrachtet, wobei ein Modell zur Lagerplatzvergabe mit DM vorgestellt wird.

2.1.2.1 Bestimmung des Lagerplatzbereiches

Als Voraussetzung für diesen Prozessschritt müssen unterschiedliche Lagerbereiche vorhanden sein. Sofern mehrere Lagerbereiche vorhanden sind, wird geprüft, ob die neu eingetroffenen Produkte zur Vervollständigung von aktiven Aufträgen benötigt werden (Backorders). Sofern Backorders vorliegen, werden diese Produkte direkt in die entsprechenden Bereiche des Lagers (Warenausgang, Versand) transportiert. Die restlichen Produkte werden in die verschiedenen Bereiche eingelagert. Dazu werden im Lagerverwaltungssystem die Transportziele der einzelnen Produkte festgelegt. Diese Entscheidungsfindung des Lagerverwaltungssystems steht im Fokus dieser Arbeit [HS08]. In der Literatur finden sich verschiedene Verfahren zur Auswahl des Lagerplatzes. Dabei existiert kein Algorithmus, welcher angibt in welchen Lagerbereich (automatisiert oder manuell) die Produkte eingelagert werden sollen. Eine Unterscheidung zwischen dem auszuwählenden Lagerbereich und dem Lagerplatz ist an dieser Stelle durchzuführen. Der Lagerbereich beinhaltet mehrere Lagerplätze, dementsprechend steht der Lagerbereich hierarchisch über dem Lagerplatz [KBW+14]. Nachfolgend werden die unterschiedlichen Algorithmen zur Lagerplatzvergabe beschrieben, welche als Grundlage für die Auswahl des Lagerbereiches verwendet werden können.

2.1.2.2 Lagerplatzvergabestrategien

Die Vergabe des Lagerplatzes steht in direktem Zusammenhang mit der Verteilung der Produkte auf den jeweiligen Lagerbereich. Grundsätzlich kann zwischen drei verschiedenen Strategien unterschieden werden [KBW+14]:

- Freie Lagerplatzvergabe (Random Slotting)
- Lagerplatzvergabe nach Kennzahlen (Slotting by Turnover Based Metrics)
- Lagerplatzvergabe nach Affinität der Produkte (Slotting by Affinity)

Die ersten beiden Strategien beinhalten jeweils nur einen Algorithmus, die Lagerplatzvergabe nach der Affinität der Produkte enthält mehrere Lösungsvorschläge. Für alle drei Strategien werden die zu Grunde liegenden Algorithmen nachfolgend erläutert.

Bei der *freien Lagerplatzvergabe*, auch Random Slotting genannt, werden die Produkte in das jeweils passende und freie Fach eingelagert. Ein Vorteil dieser Strategie stellt die Verteilung der Kommissionierer über das gesamte Lager dar. Der daraus resultierende Verkehr erstreckt sich dann ebenfalls über das gesamte Lager. Somit ist die Wahrscheinlichkeit für Lagerbereiche mit Engpässen geringer. Daraus entsteht der Nachteil, dass die Kommissionierer längere Wegzeiten in Kauf nehmen müssen. Heutzutage wird sie häufig in der Praxis angewendet und findet sich oft in der Literatur wieder [KBW+14].

Die Strategie der *Lagerplatzvergabe nach Kennzahlen* lässt sich insbesondere auf den Cube-per-order-Index zurückführen. Dieser wurde im Jahr 1963 von [Hes63] entwickelt. Als Grundlage wird die ABC-Zonung genutzt, hierbei werden häufig verwendete Produkte zusammen in einer Zone eingelagert (A-Zone) und weniger häufig verwendete Produkte in anderen Zonen (B-und C-Zone). Dieser Ansatz wurde auf die Abmessungen und das Gewicht der Produkte erweitert. Diese beiden Faktoren wurden mit in die Berechnung der Lagerplatzvergabe einbezogen und aus beiden Berechnungen wurde der endgültige Lagerplatz festgelegt. Für weitere Literatur wird auf [Hes63] und [KL76] verwiesen. [KBW+14].

Die letzte Strategie ist die *Lagerplatzvergabe nach der Affinität der Produkte*. In der Regel erhält der Kommissionierer eine Pickliste mit mehreren zu kommissionierenden Produkten. Viele Produkte auf dem Pickzettel werden häufig zusammen bestellt. Die Produkte werden als abhängig voneinander oder affin bezeichnet. Bei dieser Strategie werden die Beziehungen der Produkte untereinander betrachtet. Hintergrund der gemeinsamen Lagerung von affinen Produkten ist die Reduzierung der Wegzeit des Kommissionierers. Nicht in jedem Logistikzentrum ist das umsetzbar. Daher ist es abhängig vom Layout des Logistikzentrums, der gewählten Kommissionierstrategie und der Art der Aufträge. Die Lagerplatzvergabe nach der Affinität ist nicht immer sinnvoll: wenn viele untereinander abhängige Schnelldreher an einem Ort gelagert werden, kann es dort zu einem Stau unter den einzelnen Kommissionierern kommen. Um die Lagerplatzvergabe nach der Affinität durchzuführen werden im Folgenden einige Vorgehensweisen vorgestellt [KBW+14].

Der erste Algorithmus ‚*correlated storage* genannt, wurde von [FS89] im Jahr 1989 entwickelt. Dieser Algorithmus nimmt sich das am meisten angeforderte Produkt und sucht nach den korrelierenden Produkten. Das korrelierende Produkt wird über Häufigkeit der gemeinsamen Bestellungen von beiden Produkten bestimmt. Daraus wird zusammen mit dem meist angeforderten Produkt eine Gruppe gebildet. Zu dieser Gruppe werden so lange Produkte hinzugefügt, bis ein durch den Entscheider festgelegter Wert erreicht ist. Dieser Wert gibt an, wie groß die gebildete

Gruppe sein darf. Die jeweils hinzugefügten Produkte sind immer abhängig von dem Ausgangsprodukt. Die so entstehenden Gruppen von Produkten werden in das jeweilige Lagerverwaltungssystem implementiert, welches die Produkte nach der Affinität einlagert [FS89].

Ein weiterer Ansatz wurde von [Gar05] im Jahr 2005 entworfen. Bei diesem Algorithmus sollen *Aufträge über mehrere Lagerzonen minimiert* werden. Der Kommissionierer soll die Produkte nur aus einer Lagerzone picken und nicht aus verschiedenen Lagerzonen. Dies geschieht unter Berücksichtigung der Korrelationen von den Produkten untereinander. Für vertiefende Literatur wird [Gar05] empfohlen. Ebenfalls ein zweiphasiges Modell wurde von [KS08] entwickelt. Dieses funktioniert ähnlich zu dem von [Gar05]. Der Unterschied besteht darin, dass sie anfänglich eine *Pickfrequenz Methode* nutzen, wie den Cube-per-Order-Index. In der zweiten Phase des Modells werden dann paarweise Vertauschungen vorgenommen, damit affine Produkte näher zusammen gelagert werden können [KBW+14].

Die Entwicklung zur Optimierung der Lagerplatzvergabe hat weiterhin den Algorithmus *order oriented slotting* (OOS) von [MSH07] hervorgebracht. Dieser Algorithmus berechnet die Häufigkeit des gemeinsamen Auftretens in einer Bestellung und berücksichtigt zur gleichen Zeit die Einlagerung der Schnelldreher möglichst nah am Übergabepunkt zum nachfolgenden Bereich. Die Zielfunktion besteht aus zwei verschiedenen Berechnungen. Mit der ersten Berechnung wird die Häufigkeit der Bestellung des Produktes angegeben und mit der zweiten Berechnung wird die Anzahl der Bestellungen gezählt, die zwei gleiche Produkte enthalten. Anschließend werden die beiden Berechnungen mit der spezifischen Entfernung der Streckenführung der einzelnen Produkte multipliziert. Der Einfluss der beiden daraus entstanden Funktionen kann über einen Parameter gesteuert werden, der entweder durch den Anwender oder automatisch festgelegt werden kann [MSH07].

Eine Weiterentwicklung des OOS Algorithmus wurde unter dem Namen *pick frequency/part affinity* (PF/PA) von [KBW+10] veröffentlicht. Die Weiterentwicklung ermöglicht es, dass ein Produkt mehrmals in verschiedenen Zonen eingelagert werden kann. Weiterhin werden unabhängig vom betrachteten Zeitfenster relative Werte verwendet, um vergleichbarere Ergebnisse zu erzielen.

Die beschriebenen Algorithmen geben einen Überblick über Verfahren um die optimale Lagerplatzstrategie zu ermitteln, da jedoch keiner der beschriebenen Verfahren mit DM arbeitet, werden diese Verfahren nur als Anregung verwendet. Eine Unterstützung von DM ist bei der Bestimmung der jeweiligen Häufigkeit denkbar und wird im Verlauf der Arbeit aufgegriffen. Dies erfolgt im Rahmen der Zeitreihenprognose. Im nachfolgenden Abschnitt wird ein aktueller Artikel zum Thema Lagerplatzvergabe mit DM vorgestellt.

2.1.2.3 Lagerplatzvergabe mit Data Mining

Wie bereits im vorhergehenden Abschnitt erwähnt, wurde im Jahr 2016 von [RN16] ein Konzept zur Anwendung von DM bei der Lagerplatzvergabe vorgestellt. Das Konzept ist in Abbildung 1 dargestellt.

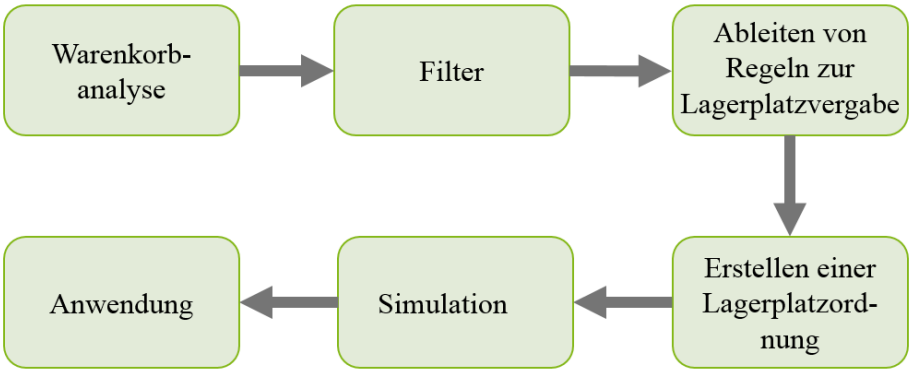


Abbildung 1: Ablaufdiagramm zur Nutzung des DM für die Lagerplatzvergabe (in Anlehnung an [RN16] S.333)

Der Ablauf enthält sechs Schritte, beginnend mit der *Warenkorbanalyse*. Die Datenbasis stellen Vergangenheitsdaten aus einem Warehouse-Management-System (WMS) dar und beinhalten unter anderem Daten über die Abmessungen, Gewicht, derzeitiger Lagerort und die Anzahl der Produkte. Mit der Warenkorbanalyse lässt sich analysieren, welche Produkte häufig zusammen bestellt werden. Genauere Erläuterungen zur Funktionsweise der Warenkorb- und Assoziationsanalyse finden sich in dieser Arbeit nicht, daher wird auf die Literatur [CL14] verwiesen. Der *Filter* dient dazu, die Ergebnisse aus der Warenkorbanalyse mit den ursprünglichen Daten zu verknüpfen und relevante Daten herauszusuchen. Im dritten Prozessschritt, dem *Ableiten von Regeln zur Lagerplatzvergabe*, werden Verfahren des maschinellen Lernens verwendet, um passende Lagerplätze für die Produkte zu finden. Die Autoren nennen in diesem Fall kein spezifisches Verfahren, jedoch verwenden sie in ihrem aufgeführten Beispiel einen Entscheidungsbaum mit dem Algorithmus ID3. Bei dem *Erstellen einer Lagerplatzordnung* werden die entwickelten Regeln aus Schritt drei verwendet, um die Produkte im Lager anzuordnen. Dieser Schritt ist abhängig von der Struktur des Unternehmens und der genauen Problemformulierung. In Schritt fünf, der *Simulation*, werden „What-If“ Szenarien erstellt, welche verschiedene Möglichkeiten der Lagerplatzvergabe simulieren. Hierbei können die Kosten herangezogen werden, die eine Entscheidungsunterstützung zur Auswahl der besten Lagerplatzvergabe darstellen. Auf Basis dieser Simulation wird die Entscheidung getroffen, in welchen Lagerbereich das Produkt eingelagert wird. Der letzte Schritt ist die *Anwendung* und Implementierung der Lagerplatzvergabestrategie [RN16].

Das entwickelte Modell ist im Bereich des DM entwickelt worden. Eine Betrachtung des gesamten Prozesses zur Nutzung von DM fehlt und ist zwingend notwendig bei der Verwendung von Daten aus einem WMS. Das entwickelte Konzept wird in dieser Arbeit nicht genauer betrachtet.

2.2 Entscheidungen

Das Ziel dieses Abschnittes ist, dem Leser einen Einblick in die wissenschaftlichen Grundlagen von Entscheidungen zu geben. Eine Entscheidung wird nach ([LGS14] S.3) folgendermaßen definiert: „Unter Entscheidung wird ganz allgemein die (mehr oder weniger bewusste) Auswahl einer von mehreren möglichen Handlungsalternativen verstanden“. Daneben sind noch zwei weitere wichtige Merkmale für eine Entscheidung hervorzuheben. Eine Entscheidung benötigt mindestens zwei Handlungsalternativen zwischen denen die Entscheidung getroffen werden muss.

Weiterhin muss eine Abweichung zwischen dem Ist-und Soll-Zustand vorliegen, welche mit der Entscheidung minimiert werden kann [GK13]. Im Folgenden wird im Kontext der Problemstellung der Prozess zur Entscheidung beschrieben.

2.2.1 Entscheidungsprozess

Eine Entscheidung durchläuft verschiedene Phasen und hat somit einen Zeitablauf. Durch diesen festen Zeitablauf wird die Entscheidung als Prozess betrachtet. Insgesamt gibt es fünf verschiedene Prozessschritte, die zu einer Entscheidungsfindung führen[LGS14]:

- Problemformulierung
- Präzisierung des Zielsystems
- Erforschung möglicher Handlungsalternativen
- Auswahl einer Alternative
- Entscheidungen in der Realisationsphase

Die verschiedenen Prozessschritte dürfen nicht getrennt voneinander betrachtet werden, weil sie voneinander abhängig sind [Til03].

Ein Entscheidungsprozess wird durch eine unbefriedigende Situation mit der Chance, eine Verbesserung zu erlangen, ausgelöst. Durch das Feststellen dieser Ist-Soll-Abweichung muss das zu lösende Problem formuliert werden. Diese *Problemformulierung* kann sofort durchgeführt werden, beispielsweise wenn eine Maschine kaputt geht, muss entschieden werden, ob sie repariert oder ersetzt werden soll. Die zweite Art der Problemformulierung benötigt einen kreativen Suchprozess im Vorfeld. Folgendes Beispiel erläutert diesen Problemfall: Ein Unternehmer möchte seinen Absatz erhöhen. Dazu muss er jedoch ein detailliertes Ziel aufnehmen, durch das er dies erreichen möchte. Es gibt mehrere Möglichkeiten dieses Ziel zu erreichen, daher steht der Unternehmer (Entscheider) in diesem Fall vor einem weiteren Entscheidungsproblem. Um eine erfolgreiche Problemformulierung durchzuführen, kann es sinnvoll sein sich weitere Informationen zu beschaffen und diese einfließen zu lassen [LGS14].

Im nächsten Prozessschritt wird eine *Präzisierung des Zielsystems* vorgenommen. Um erfolgreich entscheiden zu können werden Zielvorstellungen benötigt. An diesen Zielvorstellungen sollen die, im nächsten Prozessschritt zu entwickelnden, Handlungsalternativen beurteilt werden. Besonders in diesem Punkt ist die permanente Weiterentwicklung des Zielsystems parallel zu den anderen Schritten hervorzuheben. Zu Beginn des Entscheidungsprozesses steht eine noch sehr ungenaue Zielformulierung um den Endpunkt zu erreichen, welche sich im Laufe des Prozesses permanent weiterentwickeln und spezifizieren soll [LGS14].

Die bereits erwähnten *Handlungsmöglichkeiten* werden in diesem Prozessschritt behandelt, wobei zuerst geprüft werden muss, ob eine der Alternativen Restriktionen unterliegt. Dabei kann geprüft werden, ob beispielsweise die finanziellen Mittel überhaupt ausreichen, um das definierte Entscheidungsproblem mit dieser Alternative zu lösen. Dementsprechend sollen möglichst früh kritische Alternativen ausgeschlossen werden, um dem Entscheidungsprozess nicht noch mehr Komplexität zu verleihen. Nach dem Ausschluss der nicht umsetzbaren Handlungsalternativen erfolgt die Suche nach den umsetzbaren Handlungsalternativen. Abhängig von der Problemformulierung kann dies herausfordernd sein, dementsprechend muss der Entscheider auf Basis von Kreativität und Erfahrung Alternativen erarbeiten. Bei komplexen Entscheidungen übersteigt die Findung eines Großteils der Alternativen den Erfahrungsschatz einer Person, daher ist es sinnvoll

sich mehrere Meinungen von verschiedenen Experten einzuholen. Nach der erfolgreichen Erstellung wird versucht, das Ergebnis der entwickelten Alternativen zu prognostizieren. Der Entscheider muss die Konsequenzen seiner Alternativen abschätzen. Abhängig von der Komplexität des Entscheidungsproblems ist keine sichere Prognose möglich, da Entscheidungen teilweise bei einem unvollkommenen Informationsstand getroffen werden müssen. Dieser Informationsstand lässt sich durch wissenschaftliche Methoden verbessern, auf welche in dieser Arbeit nicht konkreter eingegangen wird [LGS14].

Darauf folgt die *Auswahl einer Alternative*. Dies ist der wichtigste Schritt im Entscheidungsprozess. Die Auswahl kann mit Hilfe verschiedener Methoden erfolgen, unter anderem auch dem DM. In dieser Arbeit wird nur diese Methode betrachtet, welche nach Tillmanns zur Entscheidungsfindung geeignet ist [Til03]; [LGS14]; [Pia10].

Der letzte Schritt ist die *Realisierungsphase*. Selbst in dieser Phase müssen abschließende Entscheidungen getroffen werden. Bei der Umsetzung einer Alternative sind Details offen geblieben und werden nun entschieden. Im Allgemeinen kann gesagt werden, dass über den gesamten Entscheidungsprozess Entscheidungen getroffen werden müssen. Diese kleinen Entscheidungen haben das Ziel das beschriebene Gesamtziel so gut wie möglich zu realisieren [LGS14].

2.2.2 Entscheidungsfindung mit Unterstützung des Data Mining

In diesem Abschnitt wird der Prozessschritt der Auswahl einer Alternative näher betrachtet, denn nur hier können entsprechende DM-Verfahren angewendet werden. In der Literatur existieren unterschiedliche Problemstellungen von Entscheidungen. Diese Problemstellungen eignen sich wiederum in unterschiedlicher Ausprägung für eine DM-Unterstützung. Im Weiteren werden diese Entscheidungsprobleme kurz erläutert und in das in dieser Arbeit vorhandene Entscheidungsproblem eingeordnet.

Insgesamt gibt es fünf verschiedene Arten von Entscheidungsproblemen. Das erste ist das wahrnehmungsdefekte Entscheidungsproblem. Dies bedeutet der Entscheider sieht noch keinen Handlungsbedarf und ihm fehlt sozusagen die „Anregungsinformation“, dementsprechend die auslösende Information für ein Entscheidungsproblem (vgl. Phase 1: Problemformulierung). Daraufhin folgt das abgrenzungsdefekte Entscheidungsproblem Hierbei sind dem Entscheider die Anzahl der Handlungsalternativen nicht vollständig bekannt. Beim wirkungsdefekten Entscheidungsproblem sind die Ergebnisse der einzelnen Handlungsalternativen abhängig von den auftretenden Umweltsituationen. Deswegen sind die Ergebnisse der einzelnen Handlungsalternativen nur schwer vorherzusagen. Im Vergleich dazu ist bei dem bewertungsdefekten Entscheidungsproblem die Bewertung des Ergebnisses vordergründig, das heißt die einzelnen Handlungsalternativen lassen sich nicht eindeutig in Anbetracht ihres Zieles bewerten. Das zielsetzungsdefekte Entscheidungsproblem stellt das letzte dar. Gibt es mehrere konfliktbehaftete Zielsetzungen in einer Entscheidungssituation und ist die Zielgröße oder das anzustrebende Zielniveau nicht bekannt, liegt das zielsetzungsdefekte Entscheidungsproblem vor [Pia10].

Von den fünf beschriebenen Arten trifft das abgrenzungsdefekte Entscheidungsproblem auf die vorliegende Ausgangsituation zu. Die Anzahl der Handlungsalternativen ist durch den Einsatz von unterschiedlichen DM-Verfahren noch nicht absehbar und es ist nicht gegeben, dass jede Handlungsalternative in Anbetracht der Ausgangssituation und der Definition von DM, Aussicht auf Erfolg hat.

Diese Einordnung erlaubt nun eine Überprüfung auf die Anwendbarkeit von DM auf die Entscheidungssituation. In der Literatur finden sich zwei verschiedene Auseinandersetzungen mit diesem Thema. Beide kommen auf unterschiedlichen Wegen zu dem Schluss, dass DM zur Entscheidungsunterstützung genutzt werden kann [Til03], [Pia10]. Im Folgenden wird nur der Ansatz von [Pia10] aus dem Jahr 2010 genauer erläutert, da dieser aktueller ist. Der Autor Piazza teilt die beschriebenen Entscheidungsprobleme unterschiedlichen DM-Verfahren zu. Die Verfahren werden im weiteren Verlauf genauer erläutert An dieser Stelle soll nur die Möglichkeit der Anwendung auf die einzelnen Entscheidungsprobleme gezeigt werden. Piazza hat zur Nutzung von DM bei Entscheidungsproblemen die Tabelle 1 entwickelt.

Tabelle 1: Nutzung der DM-Verfahren für die jeweiligen Entscheidungsprobleme (in Anlehnung an [Pia10] S.69)

	Klassifikation	Bewertung	Segmentierung	Assoziation
Wahrnehmungsdefekt	***	***	**	*
Abgrenzungsdefekt	***	***	***	*
Wirkungsdefekt	***	**	**	*
Bewertungsdefekt	***	***	-	*
Zielsetzungsdefekt	***	***	-	*

In der Tabelle steht die Anzahl der Sterne (*) für die Eignung der jeweiligen Klassen der DM-Verfahren. Das abgrenzungsdefekte Entscheidungsproblem kann nach der Tabelle mit den meisten DM-Verfahren gelöst werden. Da die anderen Entscheidungsprobleme keine Anwendung in der Arbeit finden, werden diese nicht weiter erläutert.

2.3 Herausforderungen des Data Minings in einem teilautomatisierten Logistikzentrum

Mit der Definition des Begriffes Produktvielfalt müssen die Begriffe der Produktbreite und Produkttiefe ebenfalls erläutert werden. Die Kombination aus Produktbreite und Produkttiefe ergibt die Produktvielfalt. Dementsprechend ist die Produktbreite eines Herstellers, die Anzahl an unterschiedlichen Produkten, welcher er anbietet. Die Produkttiefe wiederum beschreibt die Varianten der einzelnen Produkte [Mar16]. Die Produktvielfalt ist in den letzten Jahren exponentiell gestiegen, insbesondere in der Automobilindustrie [PH04]. In dieser Arbeit soll sich mit den Herausforderungen auseinander gesetzt werden, welche die gestiegene Produktvielfalt im Rahmen des DM mit sich bringt. Eine hohe Produktvielfalt bedeutet im Rahmen eines teilautomatisierten Logistikzentrums im Bereich des Onlinehandels viele verschiedene Produkte, welche im Lager vorrätig gehalten werden müssen. Die Verfügbarkeit eines Produktes im Lager ist ein entscheidender Aspekt bei der Kaufentscheidung des Kunden. Deswegen ist eine Reduzierung der Verfügbarkeit der Produkte für einen Onlinehändler nicht denkbar und somit eine Verringerung der Produktvielfalt aus dieser Perspektive nicht möglich [HH13]. Daher sollen die Möglichkeiten mit dem DM näher in Betracht gezogen werden. Nachdem eine Verringerung der Produktvielfalt ausgeschlossen wird, muss die Reduktion unter Anwendung von geeigneten Vorverarbeitungsschrit-

ten des DM erfolgen. Insbesondere ist dabei zu berücksichtigen, dass die Anzahl der verschiedenen Produkte in einem Logistikzentrum des Onlinehandel über das Jahr verteilt konstant ist, lediglich die Mengen der einzelnen Produkte erhöhen sich im Zeitraum November und Dezember im Vergleich zum restlichen Jahr. Dementsprechend ist die Vielfalt der einzelnen Produkte in gewisser Art und Weise begrenzt und bleibt konstant. Bei der Analyse werden nicht nur die Lagerbestandsdaten betrachtet, sondern ebenfalls die aufgegebenen Bestellungen. Die Datensätze der aufgegebenen Bestellungen sind nicht begrenzt, wie die Datensätze der Lagerbestandsdaten. Daher müssen für diese Datensätze geeignete Maßnahmen gefunden werden, um zu reduzieren, damit das DM erfolgreich angewendet werden kann. Dafür werden im folgenden Kapitel die Grundlagen gelegt, um eine erfolgreiche methodische Anwendung im Anschluss zu gewährleisten.

3 Knowledge Discovery in Databases

Der ursprüngliche Prozess, um Wissen aus Datenbanken zu gewinnen, beruhte auf manuellen Analysen und Interpretationen. Es ist jedoch ein rasantes Wachstum von Datenmengen in den untersuchten Bereichen, wie z.B.: Marketing, Finanzen usw. festzustellen. ([FPS96] S.38) schreibt dazu folgendes: „[…] this form of manual probing of a data set is slow, expensive and highly subjective". Daraus resultiert, dass für die Analyse und Interpretation großer Datenmengen in Datenbanken computergestützte Prozesse benötigt werden. Diese haben das Ziel unentdeckte und nützliche Informationen aus den Datenbanken zu gewinnen [Sha13].

Um dieses Wissen erfolgreich zu gewinnen, wird der Prozess des Knowledge in Discovery Databases (KDD) eingeführt. Dieser definiert sich nach ([FPM92] S.58) wie folgt: „KDD is the nontrivial process of identifying implicit, previously unknown and potentially useful information from data". Diese Definition aus dem Jahr 1992 wird im Jahr 1996 von ([FPS96] S. 40f.) noch weiter spezifiziert: „KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". Auf die Definition von [FPS96] wird in der heutigen Literatur weitestgehend verwiesen um KDD zu definieren. Sie wird als allgemeingültig angesehen. Der KDD Prozess besteht aus mehreren Schritten, welche iterativ zusammengehören und gegenseitig voneinander abhängen. Auf Basis diesen Prozesses und der Überführung der Datenmengen in ein kompaktes und abstraktes Modell, lassen sich durch spezifische Algorithmen Muster identifizieren [FPS96].

Im heutigen Sprachgebrauch wird DM häufig gleichgestellt mit KDD. DM ist jedoch nur ein Teilprozess des KDDs. Dennoch sind die Begriffe getrennt voneinander zu betrachten [FPS96]. Das DM stellt einen der Kernprozesse im KDD zur Wissensidentifikation dar und dient zur Identifikation und Extrahierung von Mustern aus den Datenbeständen [FPS96] [MR10]. In diesem Kapitel wird das Vorgehensmodell nach [FPS96] vorgestellt. Weiterhin wird auf die beiden Vorgehensschritte Vorverarbeitung und DM-Verfahren näher eingegangen und spezielle Verfahren in Anbetracht der Problemstellung erläutert. Es werden zwei unterschiedliche Vorgehensweisen zur Optimierung des Einlagerungsprozesses vorgestellt, weshalb in diesem Kapitel neuronale Netze zum Einsatz von Prognosen beschrieben werden. Neben den DM-Verfahren werden auch Bewertungskriterien für Klassifikationen vorgestellt. Dies ermöglicht den Vergleich von verschiedenen Ergebnissen untereinander.

3.1 Vorgehensmodell des KDD nach Fayyad et al.

Das Vorgehensmodell nach [FPS96] ist sehr weit verbreitet. Vorgestellt wurde es im Jahr 1996. Ziel ist es: „[…] die Extraktion von hochwertigem Wissen (high-level knowledge) aus Basisdaten (low-level data) durch die Anwendung einer Vielzahl von interdisziplinären Aktivitäten" ([Sha13] S.60) zu erreichen. In der Abbildung 2 ist das Vorgehensmodell zu erkennen.

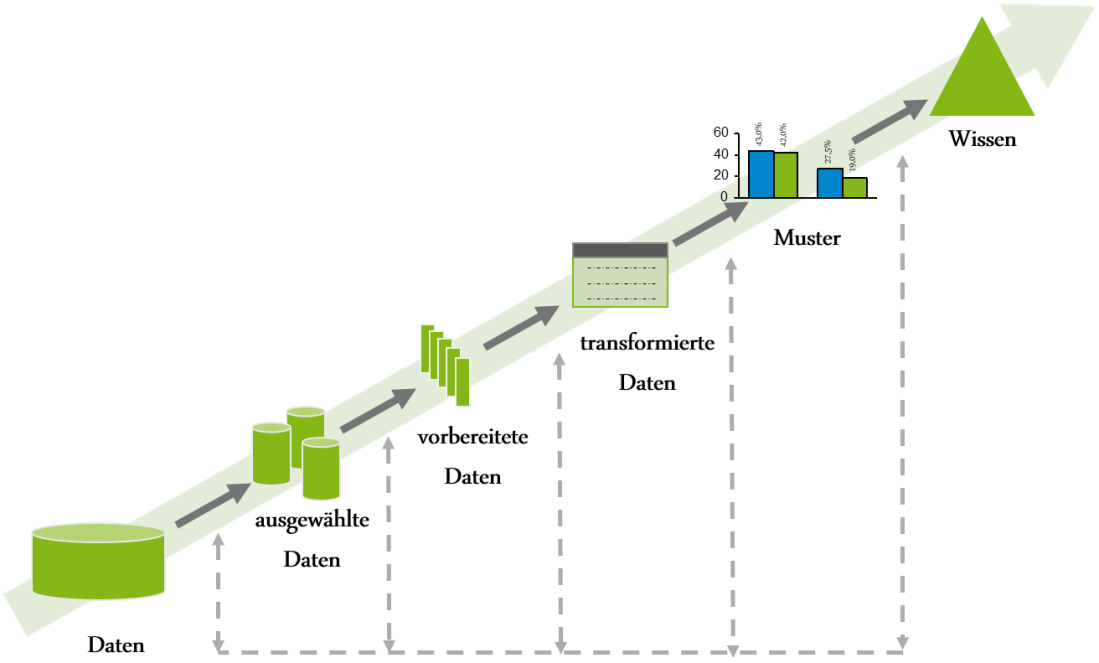


Abbildung 2: Überblick über die durchzuführenden Teilschritte (in Anlehnung an [FPS96] S.41)

Insgesamt ist der Prozess in der Abbildung 2 in sechs Schritte eingeteilt, besteht jedoch aus neun Teilschritten. Ein nicht in der Abbildung vorhandener Teilschritt ist das Domänenverständnis. Dieser Teilschritt beinhaltet neben dem Verständnis für das System auch die Zieldefinition des KDD-Prozesses. Darauf folgt die Entwicklung der Zieldaten. Diese werden dann durch verschiedene Prozesse vor- und aufbereitet. Nach der Aufbereitung der Daten werden diese transformiert und geeignete DM-Verfahren unter Berücksichtigung des definierten Ziels ausgewählt. Daraufhin werden die passenden Algorithmen und Methoden, der vorher ausgewählten DM-Verfahren angewendet, um nach Mustern zu suchen. Der letzte Teilschritt besteht darin, dass die Muster interpretiert werden und die Erkenntnisse in das System eingebracht werden können [FPS96]. Im nachfolgenden Teil werden die Teilschritte näher beschrieben, wobei sich unter dem Teilschritt DM drei Schritte wiederfinden: DM-Verfahrenswahl, Algorithmen-und Hypothesenauswahl und Mustersuche.

3.1.1 Domänenverständnis und Zieldefinition

In diesem ersten Teilschritt muss sich der Analyst der Daten ein Verständnis von dem System machen in dem er arbeiten wird. Des Weiteren müssen die Ziele der Analyse definiert werden. Das können zum Beispiel der Nachweis von Gegebenheiten sein oder die Vorhersage von Zukünftigem [Sha13]. Die Abstimmung zwischen den wirtschaftlichen und wissenschaftlichen Zielen wird heutzutage immer wichtiger. Zur genaueren Untersuchung dieses Problems hat sich ein eigenes Forschungsgebiet unter dem Namen „Domain Driven Data Mining“ entwickelt [Cao10].

3.1.2 Datenselektion

Der zweite Teilschritt besteht darin, die richtige Auswahl der Daten zu treffen. Dies ist schon ein erster Vorverarbeitungsschritt im Modell und bildet die Grundlage für eine erfolgreiche Analyse. Der Datenanalyst kann eine erste Einschätzung vornehmen, welche Daten er benötigt. Dabei

sind die benötigten Daten teilweise in verschiedenen Datenbanksystemen des Unternehmens gespeichert. Daher muss er diese gesondert ablegen und speichern. In diesem Teilschritt können ebenfalls Probleme mit der Rechnerleistung auftreten, da die zu untersuchenden Datenmengen zu groß sind. Dafür gibt es in der Literatur eine große Anzahl an Methoden um die Datenmengen zu verringern, welche im Folgenden näher erläutert werden [Sha13] [DEL14].

3.1.3 Datenvorbereitung und- bereinigung

Die Datenanalyse wird mit Realdaten aus verschiedenen Informationssystemen durchgeführt. ([GLH15] S.40) beschreibt die Realdaten folgendermaßen: „real-world data is usually incomplete, dirty and inconsistent“. Die Realdaten sind häufig nicht direkt zu verwenden und müssen erst aufbereitet werden, damit die drei Eigenschaften der Datenqualität erfüllt werden. Diese werden auf Konsistenz, Vollständigkeit und Genauigkeit geprüft [GLH15]. Auf Basis dieser Prüfung müssen die Fehler beseitigt werden, da sonst die Richtigkeit des Ergebnisses des KDD in Frage gestellt werden kann ([Sha13]. Folgende Fehler könnten hierbei auftreten [FPS96]:

- Rauschen in den Daten
- Fehlende Merkmalsausprägungen
- Messabweichungen
- Verarbeitungsfehler
- Ausreißer

Dies sind nur beispielhaft aufgeführte Fehler, welche in den folgenden Abschnitten näher erläutert werden.

3.1.4 Datentransformation

Die Datentransformation ist ein projektspezifischer Teil des KDDs. Ziel hierbei ist es die Daten so zu transformieren, dass das gesetzte Ziel aus des ersten Teilschritts erreicht werden kann [FPS96]. Das gelingt unter anderem dadurch, dass die vertikale und horizontale Dimension verringert werden kann. In diesen Teilschritt fallen ebenfalls die Aufgaben, wie Aggregation von Attributen, Umgang mit fehlenden Werten. Es findet eine Eingrenzung der Attribute statt, um etwaige Methoden zur Vorverarbeitung in den Prozess zu integrieren [Sha13].

3.1.5 Data Mining

Der Teilschritt DM unterteilt sich, wie bereits beschrieben, in drei weitere Unterpunkte. Da dies jedoch ein umfangreicher Teilschritt ist, wird er im Folgenden Verlauf noch einmal genauer mit seinen Verfahren erläutert und an dieser Stelle nur kurz erwähnt. Das Ziel des DM nach ([FPS96] S.42) beschreibt sich wie folgt: „searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering”. Das bedeutet, es werden Muster gesucht, welche sich in unterschiedlicher Form darstellen lassen. Das Ergebnis lässt sich in zwei verschiedene Bereiche einteilen: die Verifikation und die Entdeckung. Bei der Verifikation werden die vorher beschriebenen Hypothesen auf ihre Richtigkeit überprüft. Die Entdeckung wiederum findet Muster, die so nicht vorherzusagen waren [FPS96].

3.1.6 Evaluierung der Ergebnisse

Der letzte Teilschritt beschäftigt sich mit der Interpretation und Evaluation der im DM entwickelten Muster. Diese Muster werden verwendet, um Wissen aufzubauen, zu speichern und zu dokumentieren. Das Wissen kann ebenfalls auf andere Systeme übertragen oder an interessierte Personen weitergegeben werden. Eine Überprüfung mit dem vorhandenen Wissen aus der Vergangenheit ist in diesem Teilschritt ebenfalls enthalten [FPS96].

3.2 Vorverarbeitungsmethoden

Eine erfolgreiche Anwendung von DM-Verfahren benötigt eine Vorverarbeitung der einzugebenden Daten. Die Daten aus dem Warehouse Management System (WMS) oder aus anderen Datenbanken sind nicht in der Form um sie für DM-Verfahren nutzen zu können. Dies beginnt damit, dass nicht alle Daten in einer Tabelle vereint sind und erstmal alle verfügbaren Daten gesammelt werden müssen. Nach der Auswahl und dem Zusammenführen der Daten, können verschiedene Fehler in den Daten auftreten. In den jeweiligen Tabellen kann eine Vielzahl von unterschiedlichen Fehlern auftreten [CL14]. [CL14] sagen, dass die iterative Vorverarbeitungsphase im Vorgehensmodell bei bis zu 80% des gesamten Aufwandes liegt.

Ausreißer stellen bei der Datenanalyse einen klassischen Fehler dar. Ausreißer sind Werte, die nicht im normalen Wertebereich des Attributes liegen. Die Behandlung von Ausreißern sollte mit Vorsicht durchgeführt werden. Teilweise existieren in den Datensätzen ungewöhnliche jedoch korrekte Daten, welche wertvolle Informationen beinhalten können. Die Entscheidung, ob die Daten ein Ausreißer sind oder nicht, muss vom Experten getroffen werden und hängt vom konkreten Kontext ab [Run15].

Neben den Ausreißern können in den Daten fehlende, ungenaue, falsche und widersprüchliche Werte vorhanden sein. Im ersten Schritt muss analysiert werden, woher die jeweiligen fehlerhaften Daten stammen. Viele verschiedene Gründe können hinter den fehlerhaften Daten stecken. Der Anwender der Datenbank kann falsche Eingaben getätigt haben oder die Struktur der Datenbank wurde verändert, indem neue Attribute hinzugefügt wurden. Mit welchen verschiedenen Verfahren diese Fehler automatisiert beseitigt werden können, wird in den nachfolgenden Abschnitten genauer erläutert [CL14].

Die *Dimensionsreduktion* ist ein weiterer wichtiger Vorverarbeitungsschritt. Die jeweiligen Rechnerkapazitäten kommen insbesondere beim DM schnell an ihre Grenzen, daher ist eine Reduktion der Attribute häufig notwendig. Eine Reduktion kann auf zwei unterschiedliche Arten erfolgen: Attribute können einfach ausgeblendet werden oder es findet eine Aggregation von mehreren Attributen zu einem Attribut statt [CL14]. Die Aggregation von verschiedenen Attributen wird in den folgenden Abschnitten näher erläutert.

Zusammengefasst werden folgende vier Phasen betrachtet [CL14]:

- Datenselektion und-integration
- Datensäuberung
- Datenreduktion
- Datentransformation

Verschiedene Autoren nennen unterschiedliche Vorgehensweisen bei der Vorverarbeitung der Daten. Insbesondere der Schritt der Datenreduktion wird in [GLH15] nicht als notwendige Phase

beschrieben, denn eine Reduktion der Daten setzt vorverarbeitete Daten voraus. Nur durch eine Reduktion der Daten kann keine vollständige Vorverarbeitung der Daten gewährleistet werden. Deswegen sehen [GLH15] dies nur als optionalen, nicht notwendigen Schritt. In dieser Arbeit wird er als notwendiger Schritt gesehen. Aufgrund der zu großen Datenmenge werden einzelne Maßnahmen vorgestellt. Im Folgenden werden die vier genannten Prozessschritte näher erläutert und mögliche Methoden aufgezeigt.

3.2.1 Datenselektion und- integration

Bei der Datenselektion und- integration werden die notwendigen Daten zuerst ausgewählt und im darauf folgenden Schritt zusammengeführt. Die Auswahl der Daten ist die Selektion und die Zusammenführung der Daten die Integration. Ziel ist es eine Datentabelle zu erstellen mit allen notwendigen Werten. Bei der Zusammenführung können unterschiedliche Probleme auftreten. So können Redundanzen entstehen, wenn Inkonsistenzen in der Nomenklatur von Attributen vorliegen. Dabei können Attribute mit dem gleichen Namen, welche aber in unterschiedlichen Tabellen stehen, entweder redundante Informationen oder grundsätzliche verschiedene Informationen beinhalten. Dafür sollen am besten die Metadaten genutzt werden. Dort sind die Eigenschaften der Attribute beschrieben. Ebenfalls kann es vorkommen, dass es Widersprüche in der Tabelle gibt. So kann es passieren, dass für das gleiche Produkt zwei verschiedene Bezeichnungen existieren [CL14].

Mit Hilfe von zwei verschiedenen automatischen Verfahren lassen sich Redundanzen herausfinden und reduzieren. Die Verfahren unterscheiden sich in dem untersuchten Datentyp. Der Korrelationstest kann für nominelle Werte durchgeführt werden. Er vergleicht die Attribute untereinander und als Ergebnis wird eine Matrix entwickelt, in der jedes Attribut mit jedem korreliert. Hat ein Attributpaar den Wert -1 oder 1, dann korrelieren sie stark miteinander. Liegt der Wert zwischen diesem Bereich nimmt er jeweils mehr ab bis zur Mitte. Die Mitte hat den Wert 0 und Attributpaare mit diesem Wert korrelieren nicht miteinander und sind nicht redundant [GLH15]. Das zweite automatische Verfahren würde an dieser Stelle den Umfang der Arbeit überschreiten, daher wird für tiefergehende Literatur zur Datenintegration und- selektion auf [GLH15] und [Pet09] verwiesen.

3.2.2 Datensäuberung

Nachdem die Redundanzen der Daten bekannt sind, werden sie entfernt. Neben den Redundanzen sind die bereits erwähnten Fehler ebenfalls zu beseitigen. Denn diese Fehler in den Daten können zu falschen Ergebnissen bei der Anwendung von den DM-Verfahren führen. Die Säuberung der Daten kann wiederum auch ein Hinzufügen von Daten bedeuten, wenn es fehlende oder falsche Daten gibt. Dabei ist zu beachten, dass diese Daten möglichst informationsneutral gehalten werden und nicht maßgeblich das Ergebnis beeinflussen [GLH15]. Die folgenden Möglichkeiten zur Säuberung von Daten sind aus dem Buch von [CL14] entnommen.

Bei *fehlenden Daten* können verschiedene Möglichkeiten angewendet werden, um diese zu beseitigen. In Tabelle 2 sind Verfahren und deren jeweilige Anwendung aufgeführt, um fehlende Daten zu ergänzen.

Tabelle 2: Verfahren zur Säuberung von fehlenden Werten (in Anlehnung an [CL14] S.200-202; [GLH15] S.59-64)

Verfahren	Anwendung
Attribut ignorieren	• Fehlerhafte Attribute werden in Form der gesamten Spalte herausgelöscht • Bedeutet Informationsverlust, sollte daher gut überlegt sein
Werte manuell einfügen	• Fehlende Werte werden manuell eingetragen • sehr zeitintensiv und unrealistisch
Globale Konstante	• Fehlende Werte erhalten eine Unbekannte • Anwendung, wenn viele Werte fehlen oder ein leeres Feld als Information angesehen wird.
Durchschnittswerte	• Anwendbar bei numerischen Werten • Die jeweiligen fehlerhaften Daten werden mit dem Durchschnittswert des Attributes gefüllt • Einfach und häufig angewendet
Wahrscheinlichster Wert	• Fehlerhafte Werte werden durch den wahrscheinlichsten Wert ersetzt • Ermittlung über statistische Methoden, wichtig es werden ausreichend Anhaltspunkte benötigt
Häufigster Wert	• Sofern ein nichtnumerisches Attribut vorliegt, kann der häufigste Wert eingesetzt werden
Relation zwischen Attributen	• Ausnutzen von Relationen zwischen zwei Attributen • Bei numerischen Werten können mit Hilfe der Regressionsfunktion fehlende Werte berechnet werden
Datensatz als fehlerhaft kennzeichnen	• Ausschließen der Datensätze zur Weiterverarbeitung • Nur sinnvoll bei ausreichend Datensätzen

Durch das Einfügen von Werten ist es nicht möglich, dass diese Daten informationsneutral bleiben. Es ist nicht möglich mit den fehlerhaften Daten weiterzuarbeiten, da sonst die DM nicht korrekt arbeiten. Die Dokumentation der durchgeführten Veränderungen ist unabdingbar [CL14]. Neben den beschrieben Vorgehensweisen gibt es eine weitere Anzahl von Vorgehen zum Umgang mit fehlenden Daten. Dazu zählen Methoden basierend auf dem maschinellen Lernen, beispielsweise die Maxiumum-Likelihood Methode oder Verfahren basierend auf experimentell vergleichbaren Analysen. Hierzu finden sich nähere Information in Kapitel 4 von [GLH15].

Neben den fehlenden Werten können die *Daten auch verrauscht sein oder Ausreißer* haben. Um das Rauschen zu reduzieren, müssen die Daten in einer gewissen Art und Weise geglättet (angeglichen) werden. Ebenfalls müssen die Ausreißer identifiziert werden, um mögliche Maßnahmen zur Beseitigung dieser zu unternehmen. Dafür werden in Tabelle 3 verschiedene Verfahren aufgezeigt.

Tabelle 3: Verfahren zur Säuberung von verrauschten Daten und Ausreißern in Anlehnung an ([CL14] S.203-204)

Verfahren	Anwendung
Klasseneinteilung	• Gruppieren der verrauschten Daten und ersetzen durch Mittelwerte
Regression	• Beschreiben der Daten durch eine mathematische Funktion • Ersetzen der verrauschten Werte mit Hilfe von linearere Regression
Verbundbildung (clustering)	• Bilden von Clustern mit ähnlichen Werten • Ausreißer liegen dann außerhalb dieser Cluster
Kombinierte Maschine/Mensch Unterhaltung	• Computer erstellen eine Liste mit Ausreißern • Manuelle Überprüfung der Ausreißer durch den Anwender

Abschließend ist die Frage, wie Ausreißer beseitigt werden können nicht geklärt. Die beschriebenen Verfahren in der Tabelle zeigen lediglich Möglichkeiten zur Identifikation von Ausreißern. Für eine Säuberung der Ausreißer wird auf die in Tabelle 2 vorgestellten Verfahren zurückgegriffen. In der Literatur existiert eine Vielzahl an verschiedenen Arten von Rauschen und möglichen Vorschlägen zur Glättung [CL14]. Deswegen wird für vertiefende Informationen auf Kapitel 5 in [GLH15] und auf Kapitel 3 in [Run15] verwiesen.

Im ersten Teil dieses Abschnittes wurde bereits über *Inkonsistenzen und falsche Daten* hingewiesen. Dabei existieren eine Vielzahl von Fehlermöglichkeiten, welche im Abschnitt 3.2.1 bereits erläutert wurden. Bei der Vorverarbeitung der Daten kann es häufig passieren, dass die Daten nicht im definierten Wertebereich liegen oder dass sie nicht plausibel sind. In [CL14] werden als Beispiel für eine Wertbereichsverletzung, Zahlen betrachtet die lediglich einstellige natürliche Zahlen sein dürfen. Dementsprechend dürfen keine Zahlen die größer als neun sind oder kleiner als eins sind in der jeweiligen Spalte auftauchen. Ein Beispiel für nicht plausible Daten stellt folgende Problematik dar: Ein Kunde mit immer geringen Umsätzen in der Datenbank hat in der Summe einen hohen Jahresumsatz. Weiterhin können widersprüchliche Daten auftreten. Als Beispiel ist hier das Geburtsjahr, welches nicht zum Alter in der Datenbank passt. Nachdem Identifizieren solcher Probleme existieren zwei Verfahren, um diese zu beseitigen. Eine Möglichkeit besteht darin, dass der fehlerbehaftete Datensatz gelöscht wird oder bei mehreren falschen Werten die gesamte Spalte des Attributes gelöscht wird. Wobei die Zuhilfenahme von anderen Datensätzen eine zweite Möglichkeit darstellt. Es wird versucht auf der Basis von nicht fehlerhaften Werten einen plausiblen Wert zu generieren. Das Löschen von Zeilen bedeutet gleichzeitig immer einen Informationsverlust und sollte wie eingangs erwähnt wohl bedacht ausgeführt werden [CL14].

3.2.3 Datenreduktion

Im Vorverarbeitungsschritt der Datenreduktion werden die teilweise großen Datenmengen versucht zu reduzieren. Die Reduktion ist abhängig von den Daten, weil die Rechnerkapazität an ihre

Grenzen stößt. Ein weiteres Problem stellt die Interpretation im letzten Schritt des KDD-Vorgehensmodells dar, denn bei zu vielen Daten kann der Anwender kein neues Wissen entdecken. Deswegen müssen geeignete Maßnahmen gefunden werden, um die Datenmengen zu reduzieren. Insgesamt können vier verschiedene Verfahren zur Datenreduktion angewendet werden [CL14]:

- Aggregation
- Dimensionsreduktion
- Datenkompression
- Numerische Datenreduktion

Bei der *Aggregation* sollen mehrere Informationen in einem Attribut wiedergegeben werden. Hierbei kann unterschieden werden, ob eine zeilenweise Aggregation oder eine spaltenweise Aggregation vorliegt. Unter Aggregation ist auch Verdichtung zu verstehen. Dies verdeutlicht das Ziel der Aggregation von Attributen und Zeilen. Als Beispiel dienen die Umsätze einer Firma: Liegen diese monatlich vor, können diese zu einem Jahresumsatz aufsummiert werden. Nach der Berechnung ist nur noch ein Datensatz anstatt zwölf Datensätze vorhanden. Ähnlich verhält es sich mit der Aggregation von Spalten: Liegen etwa Tag, Monat und Jahr als einzelne Attribute vor, können diese zu einem Attribut Datum zusammengefasst werden. Besonders die zeilenweise Aggregation wird in dieser Arbeit angewendet [CL14].

Neben der Aggregation können die Daten auch über eine *Dimensionsreduktion* verringert werden. Dabei sollen irrelevante Daten ausgeschlossen werden. Entweder können die Daten schrittweise reduziert werden, dementsprechend immer mehr Attribute von der Gesamtmenge gelöscht werden oder die Daten werden schrittweise dem Zieldatensatz hinzugefügt und die nicht benötigten gelöscht.

In der *Datenkompression* werden die Daten wahlweise transformiert oder codiert, um somit eine Verringerung zu erzeugen. Im Vordergrund steht hierbei das Zusammenfassen von Binärattributen zu einem Byte oder aggregiert Attribute, wie bereits im Abschnitt 3.2.2 beschrieben [CL14].

Die letzte Möglichkeit Daten zu reduzieren erfolgt über die *numerische Datenreduktion*. In diesem Fall wird eine repräsentative Teilmenge von Datensätzen untersucht. Dies kann mit Hilfe von Stichproben realisiert werden. Um die Stichprobe erfolgreich auszuwählen existieren unterschiedliche Verfahren, welche im Folgenden kurz vorgestellt werden. Bei der zufälligen Stichprobe werden aus der Quelldatenmenge die Datensätze zufällig ausgewählt. Die repräsentative Stichprobe sucht ebenfalls zufällig die Daten aus der gesamten Datenmenge heraus, achtet jedoch auf die Repräsentativität der Stichprobe. Insbesondere bei Klassifikationsproblemen ist die Repräsentativität der Stichprobe zu berücksichtigen, denn jede Klasse muss mindestens einmal vertreten sein. Die Repräsentativität sollte unter der Berücksichtigung der Häufigkeitsverteilung einzelner Attribute getroffen werden. Bei der geschichteten Stichprobe werden die Datensätze zufällig ausgewählt, hier wird jedoch darauf geachtet, dass wichtige Attribute einen Wert besitzen [CL14]. In der Literatur werden noch weitere Stichproben beschrieben, jede zu erläutern würde den Rahmen der Arbeit übersteigen und findet ebenfalls im weiteren Verlauf keine Anwendung. Deswegen wird auf weitere Literatur von [GLH15] verwiesen.

3.2.4 Transformation der Daten

Der letzte Schritt der Datenvorverarbeitung beschäftigt sich mit der Transformation der Daten. Alle bis jetzt beschriebenen Vorverarbeitungsschritte können unabhängig vom gewählten DM-Verfahren angewendet werden. Dieser letzte Schritt muss immer in Abstimmung mit dem jeweiligen DM-Verfahren durchgeführt werden. Das Hauptziel dieses Schrittes besteht darin, die Daten so umzuwandeln, dass DM-Verfahren damit arbeiten können und erfolgreiche Ergebnisse liefern. Folgende Liste zeigt Beispiele, in welchen Bereichen Anpassungen vorkommen können [CL14]:

- Datentypen
- Konvertierung von Codierungen
- Zeichenketten
- Datumsangaben
- Maßeinheiten und Skalierungen

Eine Anpassung der *Datentypen* ist, abhängig vom gewählten DM-Verfahren, immer notwendig. Der Entscheidungsbaum benötigt nur nominale Werte, um ein Ergebnis zu errechnen. Im Vergleich dazu benötigt das neuronale Netz numerische Werte, um die Berechnungen durchzuführen. Daher ist es unabdingbar den Datentyp vor dem jeweiligen DM-Verfahren anzupassen. Numerische Werte können beispielsweise als nominale Intervalle dargestellt werden oder nominale Werte bekommen je nach Ausprägung einen eigenen Wert [CL14]. Eine Vielzahl von unterschiedlichen Datentypen existiert in der Literatur. Aufgrund der Komplexität werden die einzelnen Datentypen an dieser Stelle nicht näher erläutert, in der Literatur von [Pet09], [CL14] und [Run15] finden sich jedoch zahlreiche Erklärungen. In dieser Arbeit wird nur von numerischen Daten (Zahlenwerte) und nominellen (Zeichenketten, Texte) gesprochen.

Bei der Anpassung der Konvertierung von Codierungen kann es in Abhängigkeit vom jeweiligen Verfahren nötig sein, die Daten umzucodieren. Dazu zählt unter anderem die Binärcodierung, die für neuronale Netze und Assoziationsanalysen genutzt wird. Dabei werden nominale Werte als neue Attribute generiert und jedes Mal wenn das Attribut auftritt, bekommt der Datensatz den Wert 1. Wenn das Attribut nicht auftritt, hat es den Wert 0. Zu diesem Bereich zählt auch die Diskretisierung von numerischen Werten. Hierbei wird der Wertebereich von numerischen Attributen in endlich viele Teilmengen aufgeteilt. Beispielsweise kann das Alter so eingeteilt werden, dass zehn verschiedene Teilmengen entstehen. Somit kann die erste Teilmenge die Datensätze mit dem Alter von 0-10 und die zweite von 11-20 usw. beinhalten.

Die Anpassung von *Zeichenketten* beschäftigt sich mit dem Umgang von Umlauten, Groß- und Kleinschreibung und Leerzeichen in den Werten. Sofern das DM-Verfahren damit nicht umgehen kann, muss dies angepasst werden.

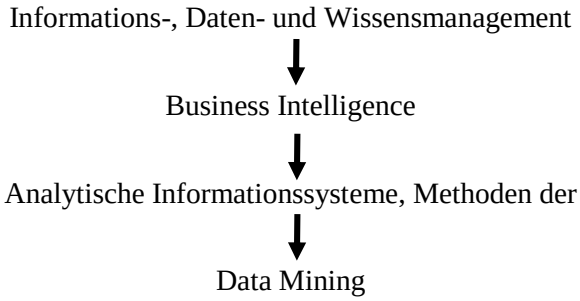
Die *Datumsangaben* müssen angepasst werden, wenn unterschiedliche Formate des Datums vorliegen. In unterschiedlichen Ländern wird das Datum anders angegeben. Datensätze aus verschiedenen Ländern müssen dementsprechend angepasst werden.

Eine weitere wichtige Möglichkeit zur Transformation stellt die *Normalisierung und Skalierung* dar. Bei der Normalisierung werden alle Werte der Attribute auf eine stetige numerische Skala transformiert. Bei der Anwendung wird meistens auf ein Intervall zwischen 0 und 1 normalisiert [CL14].

Viele weitere Verfahren wurden bereits in den Abschnitten zuvor behandelt und werden an dieser Stelle nicht noch einmal aufgegriffen. Dabei muss erwähnt werden, dass insbesondere im Schritt der Datentransformation ein iteratives Vorgehen notwendig ist, um erfolgreich DM-Verfahren anwenden zu können. Welche Art an Daten DM-Verfahren benötigen wird im Folgenden Abschnitt beschrieben.

3.3 Data-Mining-Verfahren

Dieser Abschnitt dient zur Vertiefung verschiedener DM-Verfahren. Da die DM-Verfahren den Kern des KDD bilden, wird im Folgenden eine Klassifizierung verschiedener Verfahren und eine Einordnung des DM vorgenommen. In der Literatur gibt es zahlreiche Einordnungen des Begriffs DM, in dieser Arbeit wird eine der aktuellsten Einordnungen von [CL14] vorgestellt. Die Autoren beziehen sich auf die Definition aus dem Lexikon für Wirtschaftsinformatik. In diesem wird DM als ein Bestandteil von Business Intelligence angesehen. Folgendes Schaubild nach [CL14] ordnet die Begriffe hierarchisch an:



Aus diesem Schaubild wird ersichtlich, dass DM eine Sammlung von Verfahren und Algorithmen für die Analyse von Daten ist. Daher bildet das DM eine der Grundlagen für Business Intelligence. Die Aufgaben von Business Intelligence lauten wie folgt:

- Wissensgewinnung
- Wissensverwaltung und
- Wissensverarbeitung

Überschneidungen mit dem DM sind bei der Wissensgewinnung zu erkennen. Die beiden anderen Aufgaben ermöglichen eine Konsolidierung der Ergebnisse des DM. Da die Einordnung des Begriffes DM und Business Intelligence nicht Hauptaufgabe dieser Arbeit ist, wird für vertiefende Literatur [ML13] empfohlen [CL14].

Eine erfolgreiche Anwendung von DM-Verfahren setzt Kenntnis über die vorliegende Art der Daten voraus. Insgesamt kann zwischen drei Arten von Daten unterschieden werden [CL14]. Als Beispiel für *unstrukturierte Daten* gelten Bilder oder Texte. Auf diesen Daten DM anzuwenden ist schwierig, da diese Daten vorerst in strukturierte Daten umgewandelt werden müssen [Sha13]. Ein typisches Beispiel für *semistrukturierte Daten* sind Webseiten. Diese bestehen zum Teil aus Text, weisen jedoch eine Struktur auf. Der letzte vorliegende Datentyp sind die *strukturierten Daten*, welche in dieser Arbeit vorliegen. Unter strukturierten Daten werden relationale Datenbanktabellen oder Daten in ähnlich strukturierten Dateiformaten verstanden. Die Daten sind definiert durch ihre feste Struktur, wobei die Datensätze eine feste Reihenfolge haben. Zusätzlich sind Attribute definiert und Datentypen festgelegt [CL14]; [Sha13].

Die DM-Verfahren lassen sich auf das maschinelle Lernen zurückführen. Maschinelles Lernen beschäftigt sich nach [SGS14] S.406 mit der „computergestützten Modellierung und Realisierung von Lernphänomenen“. Eine genaue Definition ist abhängig von der Definition des Begriffes Lernen. In [BK14] und [SGS14] findet sich eine vertiefende Auseinandersetzung mit dem Begriff des Lernens, welches den Umfang dieser Arbeit übersteigen würde. Grundsätzlich kann zwischen überwachtem und unüberwachtem Lernen entschieden werden.

Das *überwachte Lernen* ist die am meisten angewendete und untersuchte Art des maschinellen Lernens. Diese Aufgabe hat das Ziel, Funktionen aus Beispielen zu lernen. Ein Datensatz enthält funktionale Zusammenhänge, welche händisch oder maschinell erstellt werden. Diese Zusammenhänge sollen erkannt und in einen Algorithmus überführt werden. Bei dem *unüberwachten Lernen* existieren keine Beispiele, welche eine Gruppierung und Klassifikation vorgeben. Dementsprechend ist das Ziel des unüberwachten Lernens interessante Strukturen in unklassifizierten Daten zu finden. Beispielhaft hierfür steht die Clusteranalyse. Die Aufgabe besteht darin, Gruppen von ähnlichen Objekten zu finden und diese zu Untermengen zuzuordnen [CL14], [SGS14]. Eine Erweiterung des Begriffes stellt das *deep learning* dar. Dieser Begriff wurde in der jüngsten Zeit geprägt und verbindet das überwachte mit dem unüberwachten Lernen. Teilweise sind die DM-Verfahren des überwachten Lernens nicht für komplexe Netzwerke geeignet, hier setzt das deep learning an. Mit Hilfe der Verfahren des unüberwachten Lernens wird eine nicht exakte Lösung entwickelt, welche mit Hilfe der Verfahren des überwachten Lernens verfeinert wird. Dadurch lassen sich komplexe Netzwerke mit vielen Hierarchieebenen trainieren. Eingesetzt werden die Verfahren des deep learning bei der Handschriften- und Spracherkennung [BK14]. Für weiterführende Literatur wird [BZG+16] empfohlen.

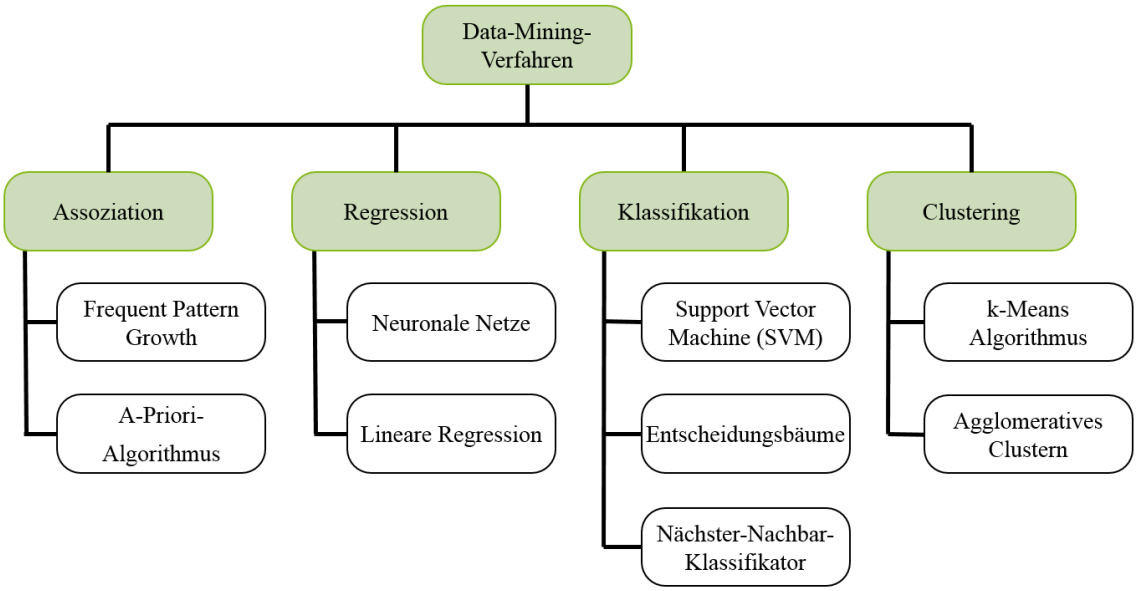


Abbildung 3: Auswahl von möglichen DM-Verfahren (in Anlehnung an [Sha13] S.69; [CL14] S.57-63; [Pet09] S.25-36)

Die DM-Verfahren lassen sich abhängig von ihrer Aufgabenstellung in verschiedene Gruppen einteilen. In dieser Auswahl existieren vier verschiedene Aufgabenstellungen mit unterschiedlichen Verfahren, wie aus Abbildung 3 zu entnehmen ist. In der Literatur sind mehr Verfahren beschrieben, die in dieser Arbeit jedoch nicht benötigt werden. Eine genaue Einteilung der

einzelnen Verfahren zu verschiedenen Gruppen gestaltet sich als schwierig. Die Verfahren des neuronalen Netzes, der Support-Vector Machine und des Nächster-Nachbar-Klassifikator sind sowohl der Regression als auch der Klassifikation zuzuordnen. Der Unterschied liegt in der Stetigkeit der Variablen. Bei der Regression werden nur stetige Variablen betrachtet, es existieren keine einzelnen Klassen [BC06]. Letztendlich wird von einer Klassifikation gesprochen, wenn die zu beschreibende Variable kategorial ist und von einer Regression, wenn die Werte für eine kontinuierlichen Variable vorhergesagt werden. Die Neuronalen Netze zählen in der Übersicht zur Regression und können ebenfalls als Klassifikationsverfahren genutzt werden. Das vorliegende Problem stellt ein Entscheidungsproblem mit zwei Varianten dar, wie in Abschnitt 2.2 bereits beschrieben. Weiterhin eignen sich für das vorliegende abgrenzungsdefekte Entscheidungsproblem drei verschiedene DM-Verfahren. Neben der Segmentierung und Bewertung zählen vor allem die Verfahren der Klassifikation dazu. Es existiert ein zweiter Ansatz zur Ermittlung des Lagerplatzes. In Abschnitt 2.1.2.2 wurden Verfahren basierend auf der Häufigkeit der Produkte beschrieben. Diese Häufigkeit soll genutzt werden, um mit Hilfe eines regressiven neuronalen Netzes den zukünftigen Absatz bestimmen zu können. Daher wird neben den Klassifikationsverfahren die Prognose mit neuronalen Netzen beschrieben.

Das Ziel der Klassifikationsverfahren besteht darin, die Daten in unterschiedliche Klassen einzuteilen. Sie gehören den Verfahren des überwachten Lernens an. Auf Basis eines Trainingsdatensatzes können automatisiert Datensätze in Klassen eingeordnet werden. Bei dem Trainingsdatensatz ist die Klasse jeweils bekannt. Als Beispiel wird die Klassifizierung der Kreditwürdigkeit herangezogen. Ein Datensatz mit Trainingsdaten und der Klasse Kreditwürdigkeit wird genutzt um in der Zukunft automatisiert, die Kunden kreditwürdig oder nicht kreditwürdig einzuschätzen. Im Folgenden werden vier verschiedene Verfahren vorgestellt, welche als Vorhersage für den zukünftigen Lagerplatz dienen und im Rahmen der Datenvorverarbeitung fehlende Werte ergänzen [CL14]. Weiterhin wird das regressive neuronale Netz zur Bestimmung der Häufigkeit näher beschrieben. Zum Abschluss des Abschnittes wird das Programm Rapidminer kurz vorgestellt, denn dies wird bei der prototypischen Umsetzung in Kapitel 5 verwendet.

3.3.1 Support-Vector-Machine (SVM)

Die Support-Vector-Machine (dt. Stützvektormethode, SVM) geht zurück auf [CV95] und ist ein geeignetes Verfahren um eine Klassifikation durchzuführen. Bei der SVM kann nur mit numerischen Attributen gearbeitet werden, wobei das Zielattribut jedoch nominal sein kann. Zur Erklärung des Verfahrens der SVM wird die Abbildung 4 genutzt. In dieser Abbildung sind Punkte in zwei verschiedenen Farben zu erkennen. Die Punkte einer Farbe entsprechen jeweils einer Klasse. Die Punkte sollen klassifiziert werden. Dafür nutzt die SVM eine Gerade, welche sie zwischen die Punkte legt (vgl. Abbildung 5).

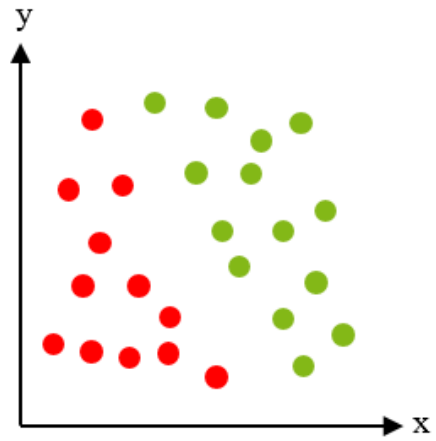


Abbildung 4: Beispiel SVM (in Anlehnung an [CL14] S.129)

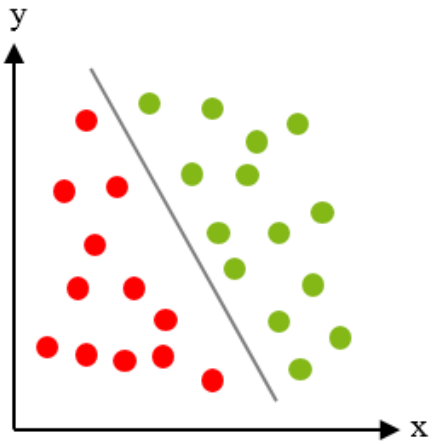


Abbildung 5: Gerade der SVM (in Anlehnung an [CL14] S.129)

Das Ziel der Klassifikation besteht darin, unbekannte Werte in Klassen einzuordnen. Mit der SVM geschieht dies durch die Zuordnung zu einer der beiden vorherigen Klassen. Bezogen auf die Abbildung 5 entweder auf die linke oder auf die rechte Seite der Geraden. Meistens besteht die Aufgabe zum Klassifizieren nicht nur aus zwei Attributen, sondern wesentlich mehr. Dafür muss die SVM im n-dimensionalen Raum verallgemeinert werden. Im Beispiel des dreidimensionalen Raumes wird anstatt einer Gerade eine Hyperebene benötigt. Bei n-Dimensionen wird eine n-1 dimensionale Ebene benötigt. Der beschriebene Abschnitt beleuchtet die Idee, welcher hinter der SVM steht. Im Folgenden werden die mathematischen Zusammenhänge erläutert und der Kernel-Trick beschrieben.

Das Ziel der SVM besteht darin eine Hyperebene zu finden, welche die Klassen am besten trennt. Dementsprechend wird die Hyperebene gesucht, welche den minimalen Abstand zu den Punkten in Abbildung 5 maximal werden lässt. Um dies zu erreichen, wird ein Stützvektor gebildet, welcher senkrecht auf der Hyperebene steht. Es ist anzumerken, dass der Fall auftreten kann, dass die Klassen nicht eindeutig durch eine Hyperebene teilbar sind. Wenn dieser Fall auftritt müssen Restriktionen verletzt werden. Um diese Verletzung in einer mathematischen Formel abzubilden wird für die Restriktion eine positive Schlupfvariable eingeführt. Das daraus entstandene Optimierungsproblem wird von der SVM berechnet. Die SVM wandelt es in ein duales Problem um und löst dieses unter zu Hilfenahme der Lagrange-Multiplikatoren und der Karush-Kuhn-Tucker-Bedingungen. Der beschriebene Ablauf ist auf linear trennbare Daten anwendbar, jedoch

existieren auch nichtlineare trennbare Daten, welche sich nicht durch eine einfache Hyperebene trennen lassen [CL14]. Um mit diesen Daten arbeiten zu können, wird der Kernel-Trick benötigt. Ziel ist es hierbei, die nichtlinearen Daten in einen höherdimensionalen Raum zu überführen, bis sich diese durch eine Hyperebene trennen lassen. Dies erfordert viel Rechenkapazität und die Rücktransformation ist in der Regel nicht brauchbar. In diesem Punkt setzen die Kernel-Funktionen an. Die Kernel-Funktion ersetzt die Transformation, indem sie die Trennfläche im mehrdimensionalen Raum mit geeigneten Funktionen beschreibt [Run15]. Zu näheren Informationen der genauen Berechnung wird auf [CV95] und [SS04] verwiesen.

3.3.2 Entscheidungsbäume

Der Entscheidungsbaum stellt eine grafische Darstellungsform von Ergebnissen einzelner Bedingungen dar. Diese einzelnen Bedingungen werden verzweigt dargestellt und aus diesen Verzweigungen können neue Verzweigungen erzeugt werden. Der Weg zur Entscheidung ist immer grafisch nachvollziehbar und liefert somit die Begründung für die Entscheidung. Zum besseren Verständnis wird mit einem Beispiel gearbeitet, welches in Abbildung 6 zu erkennen ist.

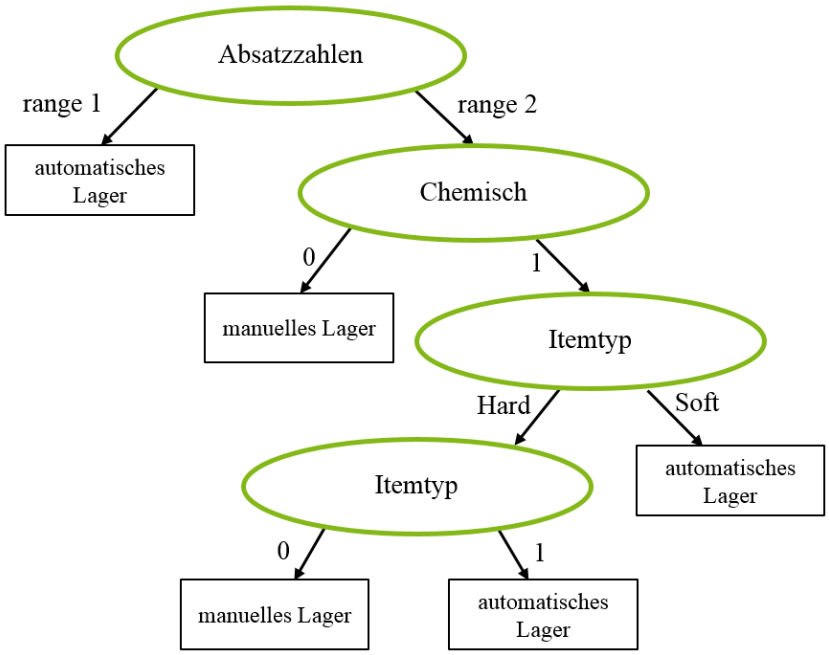


Abbildung 6: Beispiel für einen Entscheidungsbaum

In Tabelle 4 sind die dazugehörigen Datensätze aus denen der Entscheidungsbaum generiert abzulesen. Zu Beginn steht die Frage nach der Ausprägung eines Attributes, welches im dargestellten Beispiel die *Abssatzzahlen* ist. Abhängig von der Ausprägung kann die Klasse sofort vorhergesagt werden oder es muss nach weiteren Attributen gefragt werden. Die beispielhafte Entwicklung des in Abbildung 6 dargestellten Entscheidungsbaums wird im Folgenden erläutert. Ziel ist herauszufinden, wo das Produkt eingelagert wird. Der oberste Knoten heißt *Abssatzzahlen*. Ein Knoten hat jeweils eine weitere Verzweigung (in der Abbildung 6 mit grünen Kreisen dargestellt). Das Attribut hat insgesamt zwei Ausprägungen, daher entstehen zwei weitere Kanten. Die Kanten besitzen die Namen *range 1*und *range 2*. Sofern die *Absatzzahlen=range 1* sind, wird das Produkt in das automatische Lager eingelagert und die Verzweigung endet.

Tabelle 4: Daten für den Beispielentscheidungsbaum

Produkt-nummer	Chemisch	Itemtyp	Flüssig	Absatzzahlen	Lagerbereich
1	1	Hard	1	range 1	automatisches Lager
2	0	Soft	0	range 2	manuelles Lager
3	1	Soft	0	range 2	automatisches Lager
4	1	Soft	0	range 1	automatisches Lager
5	0	Hard	0	range 1	automatisches Lager
6	0	Hard	0	range 2	manuelles Lager
7	0	Soft	0	range 2	manuelles Lager
8	1	Hard	1	range 2	automatisches Lager
9	0	Soft	0	range 1	automatisches Lager
10	1	Hard	0	range 2	manuelles Lager
11	1	Hard	1	range 1	automatisches Lager
12	1	Soft	0	range 2	automatisches Lager
13	1	Hard	1	range 1	automatisches Lager
14	0	Soft	0	range 1	automatisches Lager

Dies verdeutlicht sich ebenfalls in der Tabelle, denn für alle Produkte mit der *Absatzzahl=range 1* ist der *Lagerbereich=automatisches Lager*. Ist der Baum an einer Stelle fertig verzweigt, nennt sich dies Blatt. Für die anderen beiden Attribute wird mit einer Teilmenge weiterverzweigt, bis eine eindeutige Zuordnung vorliegt. Sofern das Attribut *Absatzzahl=range2* ist, muss sich das Attribut *chemisch* angeschaut werden. Da sich diese wieder in zwei Kanten aufteilt, stellt das Attribut *chemisch* einen neuen Knoten dar [CL14].

Weiterhin gibt es eine Unterscheidung zwischen univariaten und multivariaten Entscheidungsbäumen. Bei einem univariaten Entscheidungsbaum wird an jedem Knoten genau ein Attribut abgefragt, bei multivariaten Entscheidungsbäumen können in einem Knoten mehrere Attribute genutzt werden. Multivariate Entscheidungsbäume sind wesentlich schwerer aufzubauen und dementsprechend auch schwerer zu interpretieren [CL14].

Ein weiterer Vorteil von Entscheidungsbäumen ist, dass sich durch die Baumstruktur leicht Regeln ableiten lassen. Für den in Abbildung 6 gebildeten Entscheidungsbaum könnte dies wie folgt lauten: WENN *Absatzahlen=range 2* UND *chemisch=0* DANN *Lagerbereich=manuelles Lager*. Da in dieser Arbeit die Erzeugung des Modells eines Entscheidungsbaumes im Vordergrund steht, wird die Regelableitung nicht weiter betrachtet. In dieser Arbeit wird der ID3 Algorithmus zum Lösen des Einlagerungsproblemes verwendet. Der ID3 Algorithmus geht auf [Qui86] zurück, welcher mit diesem Algorithmus unterstellt, dass der einfachste Baum die optimalste Lösung liefert. Um dies zu erreichen, arbeitet der Algorithmus mit dem Informationsgehalt eines Attributes aus dem wiederum der Informationsgewinn resultiert. Der dahinterstehende Algorithmus ist in der Literatur bereits ausführlich und verständlich beschrieben und wird an dieser Stelle nicht näher erläutert. Für weitere Informationen wird auf [CL14] S. 96-103 verwiesen.

3.3.3 Neuronale Netze

Die Neuronalen Netze (NN) sind ein Regressions-und Klassifikationsverfahren, welches in der heutigen Zeit immer mehr an Bedeutung gewinnt. Historisch beziehen sich die NN auf [MP43], welche als erstes ein logikbasiertes Modell im Bezug zu biologischen NN aufstellten. Das erste künstliche NN wurde von [Ros58] entwickelt, er nutzte das logische Modell der künstlichen NN zur Handschriftenerkennung. Eine sehr lange Zeit stagnierte die Entwicklung, in den letzten 30 Jahren hat die Entwicklung wieder rasant zugenommen. Durch die gestiegene Computerleistung und die Anwendbarkeit der NN in betriebswirtschaftlichen Problemstellungen finden sie ein immer größer werdendes Anwendungsgebiet [Cro10].

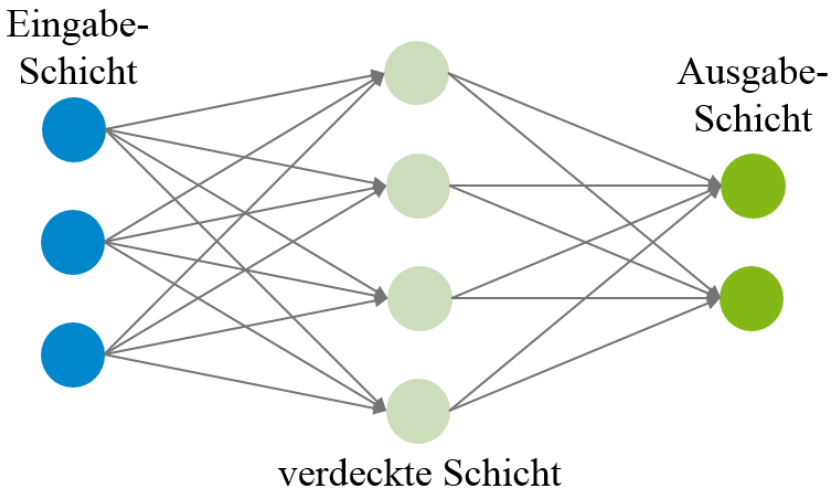


Abbildung 7: Architektur eines neuronalen Netzes (in Anlehnung an [Run15] S.70)

In dieser Arbeit werden vorwärtsgerichtete NN genutzt, diese werden ebenfalls als Backpropagation-Netze oder Mulitlayer Perceptrons bezeichnet. Bei den NN muss zuerst die Netzarchitektur bestimmt werden um im nächsten Schritt verschiedene Lernverfahren zum Trainieren der NN auf diese Architektur anzuwenden. Ein NN ist in Abbildung 7 zu erkennen, dies wird im Folgenden zur Erklärung der Funktionsweise genutzt.

Insgesamt existieren drei Bereiche in der Architektur des NN. Für ein besseres Verständnis wird das NN beginnend mit der Ausgabe-Schicht beschrieben. In der Ausgabe-Schicht sind die Ergebnisse der Klassifikation zu erkennen. Die Anzahl der Neuronen (Menge der Punkte) ist abhängig von der Codierung des Zielattributes der Klassifikation. Sofern das Zielattribut zwei unterschiedliche Ausprägungen hat, entwickelt das NN zwei Neuronen. Somit existieren für n Klassen der Ausgabe-Schicht genau n Neuronen. In der verdeckten Schicht findet die Berechnung der jeweiligen Neuronen statt. In der Eingabe-Schicht werden die zu berechnenden Daten dem NN zugeführt. Die Anzahl der Eingabe-Neuronen bestimmt sich über die Anzahl der Attribute [Pet09], [CL14]. In Abbildung 8 ist der Aufbau eines Neuronen abgebildet.

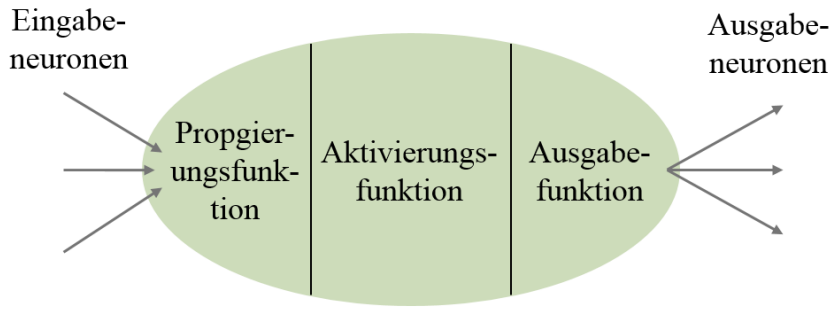


Abbildung 8: Aufbau eines Neuronen im neuronalen Netz (in Anlehnung an [Cro10] S.176)

Das Ziel der Eingabeneuronen besteht darin, die Werte aus der Außenwelt an die Propagierungsfunktion weiterzugeben. Diese Werte haben eine eigene Gewichtung, die Verbindung zwischen den einzelnen Neuronen stellen die jeweiligen Gewichte dar. Im Rahmen der Propagierungsfunktion werden die Werte der Eingabeneuronen mit den Gewichten der Verbindungen summiert, dieser Wert wird Netzeingabe genannt. Die Netzeingabe wird an die Aktivierungsfunktion weiter gegeben. Mit der Aktivierungsfunktion wird der Grad der Aktivierung des Neuron bestimmt, abhängig von diesem Zustand ist die Reaktion des Neuron bestimmbar. Erreicht die Netzeingabe einen festgelegten Schwellenwert wird das Neuron aktiviert. Sofern das Neuron aktiviert wurde, wird unter Anwendung von der Aktivierungsfunktion der genaue Grad der Aktivierung des Neurons berechnet. Die Aktivierungsfunktion ist global für die Neuronen gültig, lediglich der Schwellenwert der Netzeingabe unterscheidet sich unter den Neuronen. Im letzten Schritt berechnet die Ausgabefunktion, welche Werte an die Neuronen zu denen eine Verbindung besteht, weitergegeben werden [Cro10]. Nähere Informationen zur Aktivierungsfunktion finden sich in [Cro10] Kapitel 4 und in [SGS14] Kapitel 11.

Da der Aufbau eines Neuron und die Architektur des NN bekannt sind, wird nun ein Lernverfahren eines NN vorgestellt. Nachdem eine erfolgreiche Architektur des NN bestimmt wurde und die Eingabe-und Ausgabe-Neuronen richtig codiert wurden, kann mit dem Lernverfahren begonnen werden. Da wir in diesem Fall von einem Backpropagation-of-Error Lernverfahren ausgehen, existieren bereits Trainingsdatensätze mit dem gewünschten Ergebnis. Dementsprechend sind Fehlklassifizierungen des NN feststellbar und auf Basis dieser Fehler kann das NN sein Verhalten anpassen. Ein vorwärtsgerichtetes NN ist in Abbildung 7 zu erkennen, denn von der Eingabe-Schicht bewegen sich die Verbindungen nur in Richtung Ausgabe-Schicht. Die Hauptaufgabe des Lernens in einem NN besteht aus der Veränderung der Gewichte zwischen den Neuronen, sowie der Anpassung des bereits erwähnten Schwellenwertes. Ziel ist es, den Fehler des NN permanent zu verringern [CL14]. Eine genaue Beschreibung des Lernverfahrens findet sich bei den Entwicklern [RHW86], jedoch auch in [Run15] Kapitel 5.

3.3.4 Nutzung von Neuronalen Netzen zur Prognose von Zeitreihen

Die Nutzung von NN zur Prognose setzt Daten voraus, mit denen diese durchgeführt werden kann. Grundsätzlich kann zwischen zwei verschiedenen Arten der Prognose unterschieden werden, zum einen die Zeitreihenprognose und zum anderen die kausale Prognose. Ebenfalls existiert eine kombinierte Prognose, welche sich aus den beiden Arten zusammensetzt. Bei der Zeitreihenprognose müssen Zeitreihen vorliegen, diese definieren sich wie folgt: „Folge von Werten einer Variablen, die sich auf aufeinander folgende Zeitpunkte oder Zeiträume bezieht“ [Aue16]. Auf

Basis dieser Zeitpunkte (abhängige Variable) wird eine Prognose durchgeführt. Bei der kausalen Prognose ist der Prognosewert nicht auf die abhängige Variable zurückzuführen, sondern auf eine unabhängige Variable. Dementsprechend hängt der Prognosewert nur von externen Einflüssen ab [Cro10].

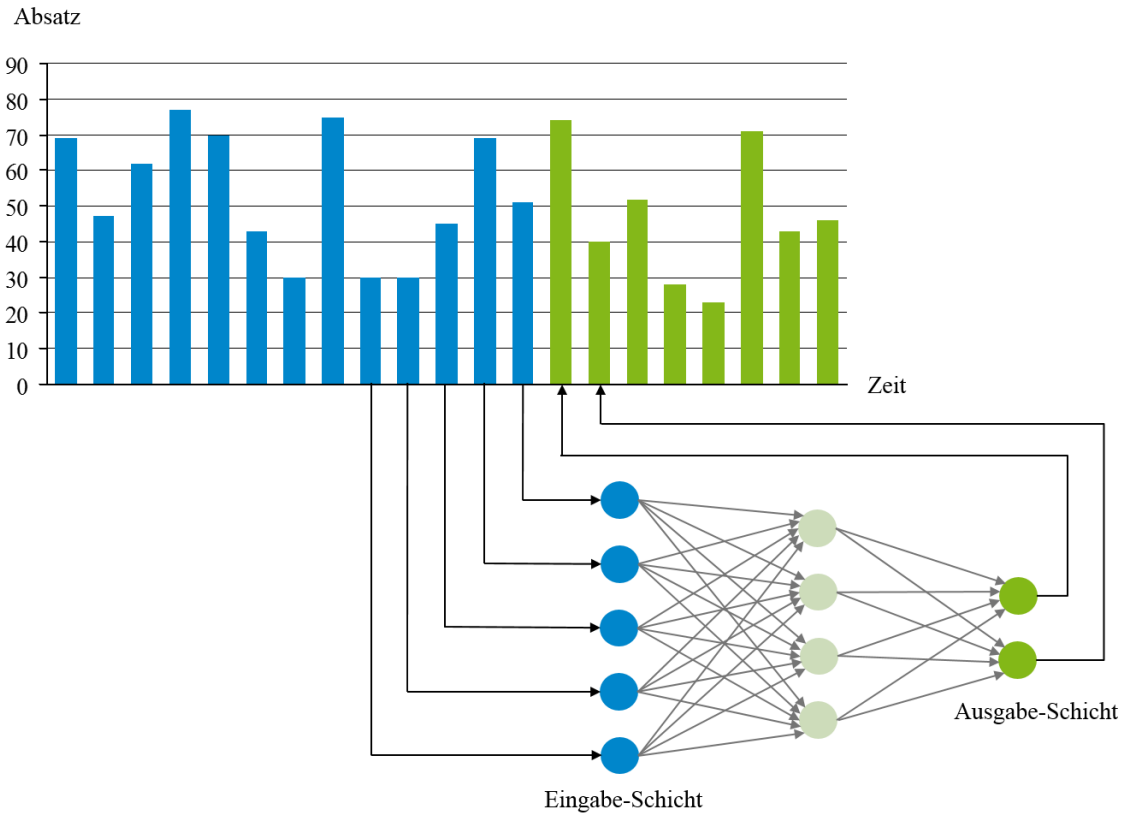


Abbildung 9: Vorwärtsgerichtetes Neuronales Netz zur Zeitreihenanalyse (in Anlehnung an [Cro10] S.228)

Sinnvoll ist es die beiden Verfahren zu kombinieren, die historischen Zeitreihen und externe Einflüsse zur Prognose zu nutzen. Dafür eigenen sich insbesondere die NN. In dieser Arbeit liegen keine verwendbaren externen Einflüsse auf die Prognose vor, daher wird die Nutzung von NN auf Zeitreihen näher erläutert.

In der Abbildung 9 sind im Säulendiagramm verkaufte Absatzzahlen über die Zeit zu erkennen. Die blauen Säulen stellen dabei die Werte der Vergangenheit dar und die grünen Säulen die prognostizierten Werte. In das NN werden in die Eingabe-Schicht die Absatzzahlen aus der Vergangenheit eingegeben. Die Größe der Ausgabe-Schicht ist abhängig vom Entscheider, dieser kann festlegen, wie weit im Voraus er die Daten haben möchte. Somit ist auch die Anzahl der prognostizierten Werte davon abhängig. Genauere Informationen finden sich hierzu in Abschnitt 3.3.3. Die abgebildete Anwendung wird ebenfalls in dieser Arbeit genutzt, es muss jedoch erwähnt werden, dass bei einer solchen komplexen Entscheidung eine kombinierte Analyse sinnvoller ist [Cro10]. Vertiefende Literatur in Bezug auf die Nutzung von NN zur Prognose befindet sich in [Cro10].

Die benötigte Anzahl an Vergangenheitsdaten um eine erfolgreiche Prognose zu gewährleisten, wird in der Literatur nicht näher spezifiziert. Die Größe des Prognosezeitraumes ist ebenfalls abhängig vom jeweiligen Prognosegegenstand und der Entscheidungssituation [Gra80]. In die-

sem Fall liegt ein Problem mit der Lagerung vor. Dementsprechend müssen die jeweiligen Absatzzahlen prognostiziert werden um die Anzahl der Produkte zu bestimmen, welche ausgelagert werden. Bei hohen Absatzzahlen werden die Produkte in den automatisierten Teil eingelagert, bei niedrigen Absatzzahlen in den manuellen Teil. In dieser Arbeit wird auf diese Problematik und das daraus resultierende Optimierungsproblem noch genau eingegangen.

3.3.5 Funktionsweise von Rapidminer

Rapidminer stellt ein Programm dar, welches im Rahmen von DM-Analysen genutzt wird. Dabei liegen die Anwendungsmöglichkeiten weit verbreitet, es kann in unternehmerischen Bereichen oder auch in der Forschung genutzt werden. Die Benutzerfreundlichkeit von Rapidminer ist hoch, denn die zu verwendenden Operatoren müssen nicht programmiert werden, sondern können per Drag&Drop in den Arbeitsbereich gelegt werden. Dies ermöglicht eine einfache Anwendung von Vorgehensmodellen ohne vorherige Programmierkenntnisse. Neben Rapidminer existieren weitere Produkte zur Durchführung von DM-Verfahren, Rapidminer stellt jedoch die meisten Operatoren und wird aus diesem Grund in dieser Arbeit verwendet. Ein weiterer Vorteil besteht darin, das Rapidminer eine Vielzahl an unterschiedlichen Dateiformaten lesen kann bis zu dem direkten Lesen der Daten aus Datenbanken. Rapidminer ist weiterhin in Java geschrieben und auf einer Vielzahl von Plattformen einsetzbar. In den nachfolgenden Abschnitten wird jeweils nur der Arbeitsbereich des Programmes gezeigt, auf eine weitere Beschreibung des Programms wird verzichtet, da viele Tutorials und Hilfestellungen im Internet zu finden sind.

3.4 Verfahren zur Messung von Klassifikationsergebnissen

Durch die Anwendung von mehreren DM-Verfahren stellt sich die Frage nach Möglichkeiten, um die Verfahren miteinander zu vergleichen. Das Problem besteht insbesondere darin, das jedes Verfahren auf andere Art und Weise seine Ergebnisse optimiert. Das NN optimiert über den auftretenden Fehler und der Entscheidungsbaum über den Informationsgehalt. Deswegen ist eine Gegenüberstellung der einzelnen Verfahren nicht exakt möglich, dennoch in dieser Arbeit notwendig. Daher werden im Folgenden drei verschiedene Möglichkeiten vorgestellt. In der Literatur existieren noch weit mehr Verfahren, die Folgenden sind die Bekanntesten zum Vergleich von DM-Verfahren.

3.4.1 Trefferwahrscheinlichkeit

Die Trefferwahrscheinlichkeit wird in der Literatur auch Erfolgsrate (engl. accuracy) bezeichnet. Um die Erfolgsrate berechnen zu können, muss vorher die Fehlerrate errechnet werden. Die Fehlerrate gibt an, wie hoch der relative Anteil der falsch klassifizierten Beispiele der Gesamtmenge ist. Sie wird durch folgende Formel beschrieben:

$$Fehlerrate = \frac{Falsche\ Klassenzuordnung}{Alle\ Klassenzuordnung}$$

Ebenfalls ist es möglich die Fehlerrate bei einer numerischen Vorhersage zu treffen, da wir aber nur nominale Vorhersagen treffen, wird dies vernachlässigt. Die Erfolgsrate berechnet sich wie folgt:

$$Erfolgsrate = 1 - Fehlerrate$$

Die Erfolgsrate dient dazu, dem Anwender ein subjektiv besseres Ergebnis zu präsentieren, weil Werte im hohen zweistelligen Prozentbereich aufgezeigt werden. Die beschriebenen Werte sind jeweils Prozentangaben [CL14].

3.4.2 Kennwerte zur Evaluierung der Klassifikation

Neben der bereits beschriebenen Trefferwahrscheinlichkeit, gibt es weitere Möglichkeiten das Klassifikationsergebnis zu messen. Insgesamt kann zwischen vier verschiedenen Fällen richtiger und falscher Klassifikation unterschieden werden. Im folgenden Beispiel wird davon ausgegangen, dass der Klassifikator gute Kunden vorhersagen soll [Run15], [CL14]:

- TP (richtig positiv) – Ein guter Kunde wird als ein guter erkannt
- TN (richtig negativ) – Ein nicht guter Kunde wird als nicht guter erkannt
- FP (falsch posistiv) – Ein nicht guter Kunde wird als guter erkannt
- FN (falsch negativ) – Ein guter Kunde wird als nicht guter erkannt

Mit Hilfe von diesen vier Einordnungen der Klassifikation kann eine Reihe von weiteren Kennwerten berechnet werden, welche in Tabelle 5 dargestellt sind.

Tabelle 5: Berechnung von Kennwerten zur Messung des Klassifikationsergebnis ([CL14] S.228-229)

Name des Kennwertes	Berechnung	Beschreibung
Korrekte Klassifikation	$T = TP + TN$	Alle korrekten Vorhersagen
Falsche Klassifikation	$F = FP + FN$	Alle falschen Vorhersagen
Relevanz	$R = TP + FN$	Anzahl der guten Kunden
Irrelevanz	$I = FP + TN$	Anzahl der nicht guten Kunden
Positvität	$P = TP + FP$	Anzahl der als gut klassifizierten Kunden
Negativität	$N = TN + FN$	Anzahl der als nicht gut klassifizierten Kunden
Korrektheitsrate	$KH = \frac{T}{n}$	Anteil der korrekt klassifizierten Kunden
Inkorrektheitsrate	$IKH = \frac{F}{n}$	Anteil der nicht korrekt klassifizierten Kunden
Richtig-positiv-Rate	$TPR = \frac{TP}{R}$	Wie oft wurde ein guter Kunde als solcher klassifiziert (Sensitivität, Revall, Trefferquote)
Richtig-negativ-Rate	$TNR = \frac{TN}{I}$	Wie oft wurde ein nicht guter Kunde auch als solcher klassifiziert?
Falsch-positiv-Rate	$FPR = \frac{FP}{I}$	Wie oft wurde ein nicht guter Kunde als guter klassifiziert?
Falsch-negativ-Rate	$FNR = \frac{FN}{R}$	Wie oft wurde ein guter Kunde als nicht guter klassifiziert?
Positiver Vorhersagewert	$precision = \frac{TP}{P}$	Wie oft ist ein als gut vorhergesagter Kunde ein guter Kunde?

Negativer Vorhersagewert	$\frac{TN}{N}$	Wie oft ist ein als nicht gut vorhergesagter Kunde ein nicht guter Kunde?
Negative Falschklassifikationsrate	$\frac{FN}{N}$	Wie oft ist ein als nicht gut vorhergesagter Kunde ein guter Kunde?
Positive Falschklassifikationsrate	$\frac{FP}{P}$	Wie oft ist ein als gut vorhergesagter Kunde ein nicht guter Kunde?

Von den Kennwerten in Tabelle 5 werden meistens mehrere verwendet, um das Klassifikationsergebnis zu messen. Um beispielsweise ein PR-Diagramm zu erzeugen, wird der Kennwert Precision mit dem Kennwert Recall in einem Diagramm dargestellt. Eine weitere Möglichkeit ergibt sich mit den Kennwerten Richtig-positiv-Rate (TPR) und der Falsch-postitiv-Rate (FPR). Daraus lässt sich das Receiver Operating Characteristic (ROC) oder auch Area under the Curve (AUC) genannt, darstellen. Dieses wird im Folgenden Abschnitt näher beschrieben.

3.4.3 Receiver Operating Characteristic

Beim ROC-Diagramm wird auf der Ordinate die TPR aufgetragen und auf der Abszisse die FPR. Das ROC-Diagramm ist nur auf Klassifikationen anzuwenden, welche als Ergebnis zwei Ausprägungen des Zielattributes besitzt. In der Abbildung 10 ist ein solches ROC-Diagramm zu erkennen.

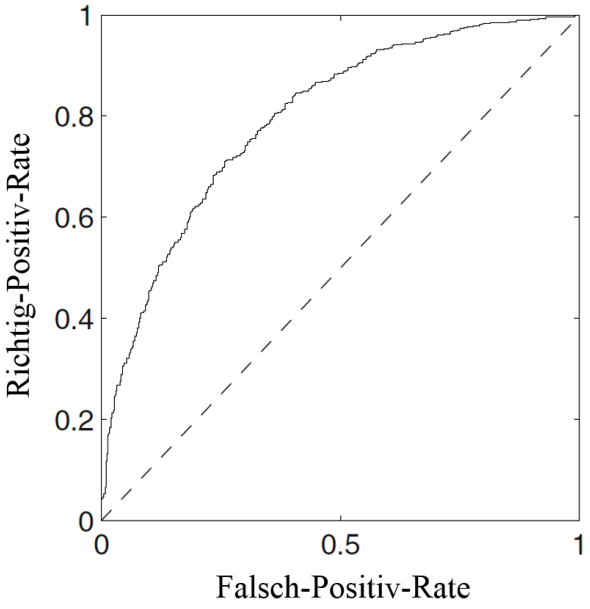


Abbildung 10: ROC-Diagramm (in Anlehnung an [Run15] S.88)

Die Güte einer Klassifikation lässt sich im ROC-Diagramm als ein Punkt darstellen, welche einen gegebenen Klassifikator, sowie einen gegebenen Datensatz voraussetzt. Sofern das gewählte Validierungsverfahren geeignet ist, lassen sich TPR und FPR auf der Validierungsmenge als typische Klassifikationsgüte des verwendeten Klassifikators interpretieren. Dies ist mit fast allen Klassifizierungsverfahren möglich, daher ist die Klassifikationsgüte unterschiedlicher Verfahren gut vergleichbar mit ROC-Diagrammen. Dementsprechend hat ein idealer Klassifikator 100% TPR und 0% FPR. Dieser Klassifikator würde sich in der oberen linken Ecke (0;1) des Diagramms befinden, dementsprechend sollten gute, jedoch nicht-ideale Klassifikatoren sich möglichst nah

an der oberen linken Ecke aufhalten. Bei der oberen rechten Ecke (1;1) liefert der Klassifikator immer ein positives Ergebnis, also 100% FÜR und 100% TPR. Im Gegensatz dazu liegt in der unteren linken Ecke (0;0) immer ein negatives Ergebnis vor, das bedeutet 0% von beiden Kennwerten. Die untere rechte Ecke (1;0) beschreibt ein Klassifikator der immer das falsche Ergebnis liefert. Dieser Klassifikator lässt sich invertieren, das heißt aus einer positiven Klassifikation wird eine negative und umgekehrt. Damit kann die Klassifikationsgüte an der Hauptdiagonalen (gestrichelte Linie) gespiegelt werden. Aufgrund dessen werden nur Klassifikatoren betrachtet, deren Güte über der Hauptdiagonalen liegen. Hat der Klassifikator keine Entscheidungsaussagen, würde die ROC-Kurve direkt auf der Diagonalen liegen [Run15]. Ein weiterer abzulesender Kennwert ist die Area under the Curve (AUC). Dieser Kennwert beschreibt die Fläche unter der ROC-Kurve (durchgezogene Linie). Sie ist das Maß für die Qualität des Klassifikators, je größer die Fläche ist, desto besser ist der Klassifikation und desto höher ist der AUC-Wert. Ebenfalls kann der AUC-Wert als Wahrscheinlichkeit interpretiert werden, dass ein positiver Wert tatsächlich auch als solcher klassifiziert wird. Der Wert für AUC liegt im Intervall von 0,5 (nutzlose Klassifikation) bis 1,0 (perfekte Klassifikation) [HM82]. Für vertiefende Literatur wird auf [HM82] verwiesen.

4 Entwicklung des KDD-Vorgehensmodells zur Optimierung des Einlagerungsprozesses

Dieses Kapitel hat das Ziel ein KDD-Vorgehensmodell zur Optimierung des Einlagerungsprozesses zu entwickeln. Dafür wird im ersten Abschnitt eine genaue Problemstellung formuliert. Auf Basis dieser Problemstellung werden die Prozesse in einem teilautomatisierten Logistikzentrum beschrieben und die Datenbankstruktur zur Steuerung des Systems näher erläutert. Die in den Datenbanken vorhandenen Daten werden mit ihren Attributen vorgestellt. Im ersten Abschnitt werden die Grundlagen für ein Verständnis des Systems gelegt, daher ist er für beide Handlungsalternativen gleich. Dabei wird an manchen Stellen Bezug auf Beispieldaten genommen, dies dient einem besseren Verständnis des Vorgehensmodells.

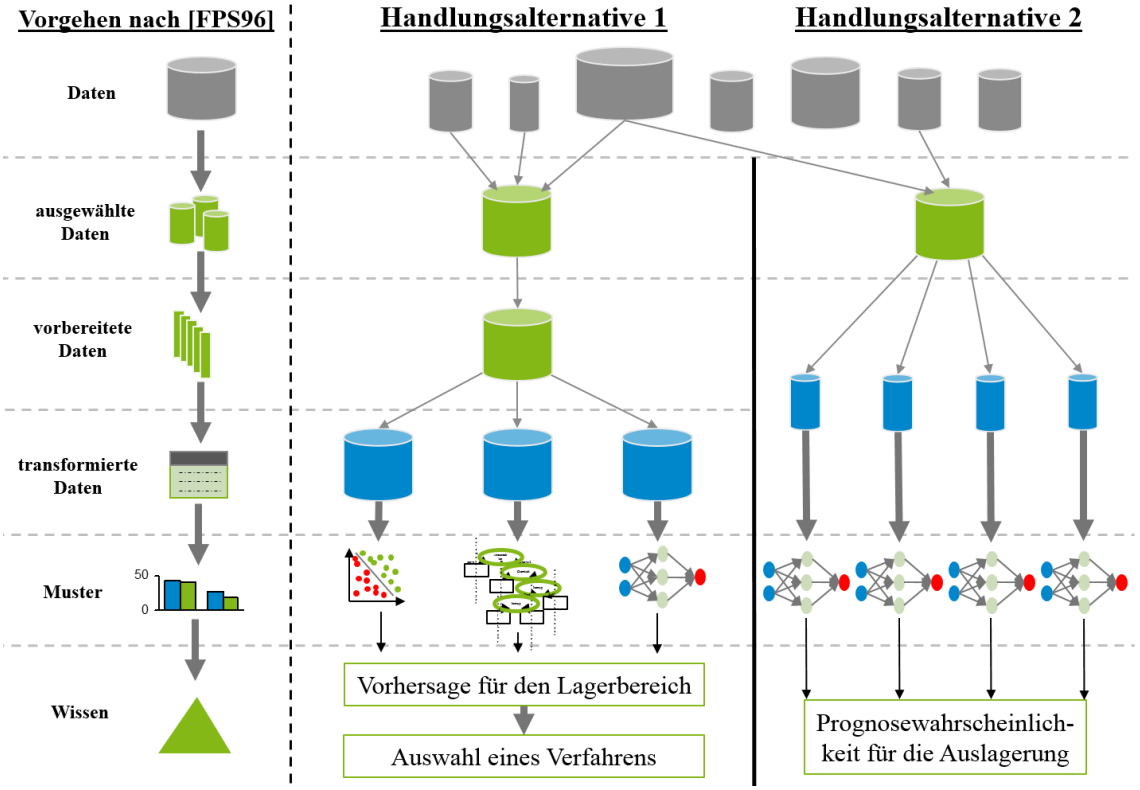


Abbildung 11: Vorgehensweise in Kapitel 4

Die Abbildung 11 zeigt drei verschiedene Vorgehensweisen und zugleich den Aufbau dieses Kapitels. Dabei stellen die zylindrischen Formen in grau, grün und blau symbolhaft die jeweiligen Tabellen dar, welche in den jeweiligen Vorgehensmodellen genutzt werden. Die unterschiedliche Farbe zeigt eine Veränderung vom Ausgangszustand an. Diese Notation wird im kompletten Kapitel 4 beibehalten. Eine von den Vorgehensweisen stellt das KDD-Vorgehensmodell nach [FPS96] dar, welches in Abschnitt 3.1 bereits beschrieben wurde. Ausgehend von diesem Vorgehensmodell soll ein Neues entworfen werden, welches den Prozess der Einlagerung optimiert. Dafür wurden zwei Handlungsalternativen gewählt, welche sich in der Abbildung 11 wiederfinden. Die Motivation zur Entwicklung der beiden Handlungsalternativen lässt sich auf zwei verschiedene Einflüsse zurückführen. In der Literatur herrschen zwei verschiedene Vorgehensweisen und es soll die Möglichkeit der Anwendung von DM untersucht werden. Dabei bezieht sich

die erste Handlungsalternative auf das Vorgehen in Abschnitt 2.1.2.3, eine Lösung zur Bestimmung des Lagerbereiches mit Hilfe von DM. Die zweite Handlungsalternative beruht auf Abschnitt 2.1.2.2, dabei handelt es sich um die Lagerplatzvergabestrategien. Bei dieser Handlungsalternative soll der Parameter zur Bestimmung der Häufigkeit des Produktes mit Hilfe von DM optimiert werden. Der zweite Einfluss stellt das praktische Problem dar, bei der Handlungsalternative Eins wird untersucht, ob sich das Problem nur mit dem Einsatz von DM lösen lässt und eine direkte Bestimmung des Lagerbereichs möglich ist. Bei der zweiten Handlungsalternative wird ein Parameter in einem bereits vorhandenen Algorithmus mit Hilfe von DM optimiert.

Die erste Handlungsalternative verfolgt das Ziel eine Vorhersage für den Lagerplatzbereich auf Basis der derzeitigen Information zu treffen. Dabei durchläuft sie verschiedene Phasen der Datenvorverarbeitung. Bei der Transformation der Daten werden diese auf die jeweilig anschließenden DM-Verfahren angepasst. Mit den angepassten Daten werden drei DM-Verfahren durchgeführt und diese liefern als Ergebnis eine Vorhersage für den Lagerbereich. Die Ergebnisse der Verfahren werden mit verschiedenen Kennwerten verglichen und eines ausgewählt.

Bei der zweiten Handlungsalternative wird auf die gleiche Datenbasis zugegriffen. Dabei wird das Ziel verfolgt eine Prognosewahrscheinlichkeit für die Auslagerung des Produktes auf Basis von Zeitreihen für die Zukunft zu treffen. Dafür werden zwei Vorverarbeitungsmaßnahmen durchgeführt, um für jedes Produkt die Daten für die in der Vergangenheit getätigten Auslagerungen zu ermitteln. Für jedes Produkt muss eine eigene Tabelle angelegt werden, um das DM anwenden zu können. In diesem Punkt wird überlegt, ob nicht Produkte zusammengefasst werden können, um die Dimension zu reduzieren. Für jedes Produkt im Logistikzentrum wird ein NN trainiert, welches daraufhin eine tageweise Prognose für einen bestimmten Prognosezeitraum generiert. Zur Bestimmung des Lagerbereiches muss dieser Wert noch in einen Algorithmus eingepflegt werden. Dabei kann der Algorithmus ein neu generierter sein, oder aus dem Abschnitt 2.1.2.2 stammen.

Den Abschluss des Kapitels stellt ein Vergleich der beiden Vorgehensweisen dar, um das geeignetste Vorgehensmodell zu definieren. Insbesondere wird über die Transparenz und Implementierung der beiden Vorgehensweisen diskutiert. Das Kapitel beginnt mit der Problemstellung und Erklärung der Prozesse in einem teilautomatisierten Logistikzentrum.

4.1 Problemstellung und Domänenverständnis

Das Ziel dieser Arbeit besteht darin, den derzeitigen Einlagerungsprozess zu optimieren. Dafür wird im Folgenden eine genaue Problemstellung vorgestellt und auf dieser werden die beiden Handlungsalternativen festgelegt. Das zugrunde liegende Problem besteht aus der Entscheidung, ob das Produkt in den automatisierten oder in den manuellen Bereich des Logistikzentrums eingelagert werden soll. Die Lagerplatzvergabe in den beiden Bereichen wird in dieser Arbeit keine Betrachtung finden, es wird die Zuordnung zu einem der beiden Bereiche bestimmt. Derzeit ist die Anzahl der Produkte, welche in den automatisierten Bereich eingelagert werden zu niedrig. Diese Anzahl soll erhöht werden, damit die Anzahl der Produkte verringert werden kann, welche in den manuellen Bereich eingelagert werden. In Abbildung 12 wird die Problematik auf Basis der Beispieldaten veranschaulicht, insbesondere wird ein Fokus auf die beiden unterschiedlichen Phasen gelegt, wie in Abschnitt 2.3 bereits beschrieben.

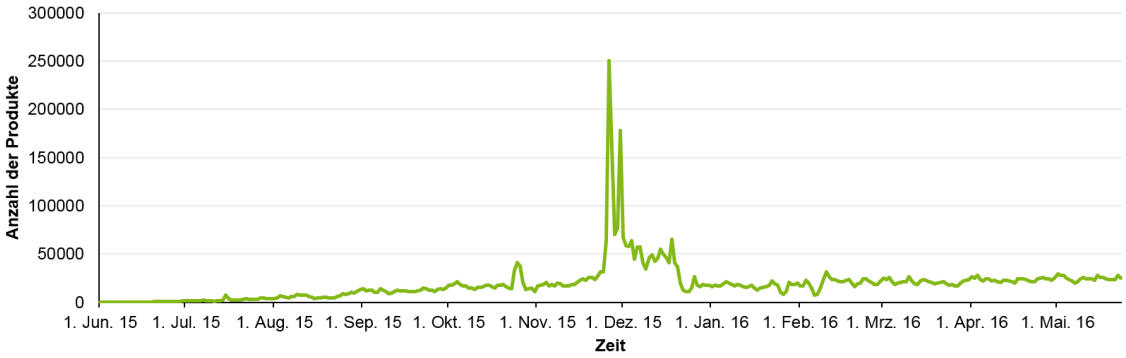


Abbildung 12: Anzahl der verschickten Produkte in einem Jahr

Die Abbildung 12 zeigt die Verteilung der Anzahl der verschickten Produkte über das Jahr und somit die Auslastung des Logistikzentrums über das Jahr. Dabei ist zu berücksichtigen, dass der Zeithorizont der Daten ungefähr ein Jahr zurück reicht. Ein eindeutiger Verlauf liegt in den Monaten Juni – Mitte November und Ende Dezember bis Mai vor. In diesem Zeitraum hat das Logistikzentrum eine normale Auslastung. Die Anzahl der verschickten Produkte liegt immer unter 50.000 pro Tag. Aufgrund der niedrigen Anzahl an benötigten Produkte, sollen in diesem Zeitraum so viele Produkte wie möglich in den automatisierten Teil eingelagert werden. Aus dem automatisierten Teil können über 144.000 Produkte am Tag ausgelagert werden. Bei der Betrachtung auf den Zeitraum zwischen Mitte November und Ende Dezember wird dieser Wert um mehr als das Doppelte überschritten. Daher muss in diesem Zeitraum der manuelle Teil mehr in den Fokus rücken, durch den Einsatz von mehr Arbeitskräften kann eine Erhöhung des Durchsatzes aus dem manuellen Teil erreicht werden und somit mehr Produkte bereitgestellt werden. Letztendlich muss diese Unterscheidung in die beiden Phasen bei der Betrachtung der Problemstellung berücksichtigt werden.

Die Zuordnung der Produkte zu dem jeweiligen Bereich ist abhängig von unterschiedlichen Restriktionen. Im automatisierten Bereich können über einen Zeitraum nur eine bestimmte Anzahl an Kisten ausgelagert werden. Somit können nicht alle Produkte in den automatisierten Bereich eingelagert werden, sondern es muss eine Aufteilung erfolgen. Weiterhin sind die Lagerplätze im automatisierten Bereich von ihrer Größe beschränkt, dementsprechend fallen zu große Produkte für die Einlagerung in den automatisierten Bereich heraus. Eine gesamte Anzahl an Restriktionen wird an anderer Stelle näher betrachtet.

Es lässt sich feststellen, dass bei der Optimierung eine Reihe von Unbekannten aufkommt. Diese sollen mit den beiden Handlungsalternativen gelöst werden. Die beiden Handlungsalternativen sind aus dem Abschnitt 2.1.2 hervorgegangen. Für die Bestimmung des Lagerbereiches unter Anwendung von DM, wurde bereits ein Vorgehensmodell vorgestellt, dies ist aufgrund der Komplexität der Daten nicht anwendbar. Die zweite Vorgehensmöglichkeit bestimmt die Häufigkeit des Produktes in Bezug auf die Auslagerung. Diese Häufigkeit lässt sich mit DM bestimmen, um eine endgültige Entscheidung für den jeweiligen Lagerbereich zu treffen, muss diese Häufigkeit in einem Algorithmus integriert werden. Für ein besseres Verständnis des Systems und der Problemstellung werden die Prozesse in einem teilautomatisierten Logistikzentrum im Folgenden erläutert. Sofern das Domänenverständnis von den folgenden Parametern abweicht, ist die Anwendung des entwickelten Vorgehensmodells vorher genau zu prüfen.

Ebenfalls muss festgelegt werden, wie häufig die Zuordnung zu den einzelnen Lagerbereichen stattfindet. Daher muss für die beiden Handlungsalternativen ein Intervall festgelegt werden, wie häufig das jeweilige DM-Verfahren durchgeführt werden muss. Dieses Intervall dient dazu, eine möglichst effiziente Zuordnung treffen zu können und ist ebenfalls abhängig von externen Faktoren. Eine genaue Auseinandersetzung mit dieser Problematik erfolgt im Vergleich der einzelnen Handlungsalternativen.

4.1.1 Prozesse in einem teilautomatisierten Logistikzentrum

In diesem Abschnitt werden die zu Grunde liegenden Prozesse eines teilautomatisierten Logistikzentrums beschrieben. Eine Aufteilung der Prozesse erfolgt zwischen dem Materialfluss und dem Informationsfluss. Zum besseren Verständnis der Prozesse wird eine schematische Abbildung des Aufbaues eines teilautomatisierten Logistikzentrums genutzt.

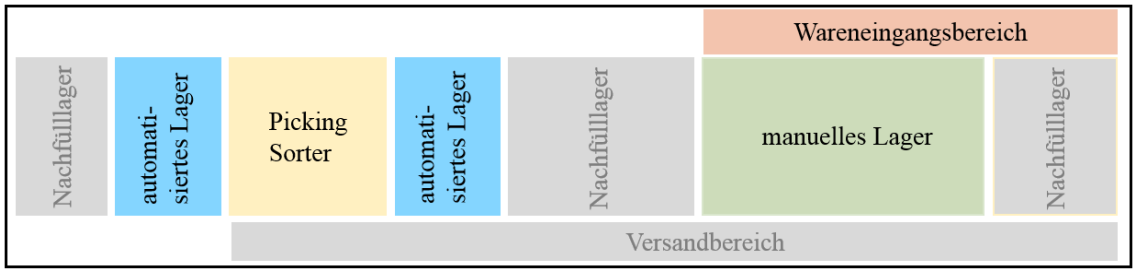


Abbildung 13: Aufbau eines teilautomatisierten Logistikzentrums

In Abbildung 13 ist ein beispielhafter Aufbau eines teilautomatisierten Logistikzentrums. Die Abbildung ist stark vereinfacht, der detaillierte Aufbau findet sich im Anhang A1. Es sind verschiedenfarbige Bereiche zu erkennen, wobei die grauen Bereiche nur indirekt mit der Problemstellung in Berührung kommen. Das Nachfülllager hat keine Auswirkungen auf die Einlagerung der Produkte, weil die Zuordnung des Lagerbereiches schon festgelegt wurde. Im Versandbereich werden keine Entscheidungen im Zusammenhang mit der Einlagerung getroffen, weil dieser den letzten Bereich in einem Logistikzentrum darstellt. Die beiden Bereiche werden im Folgenden nicht weiter betrachtet.

Der *Wareneingangsbereich* ist in der Abbildung 13 rot markiert und ist der Ausgangspunkt für alle folgenden Prozesse. In diesem Bereich werden die Produkte vereinnahmt und den entsprechenden nachfolgenden Bereichen zugeordnet, somit wird die Entscheidung zur Einlagerung in diesem Bereich getroffen.

Der *automatisierte Teil* ist in der Abbildung 13 blau markiert. Insgesamt existieren davon zwei gleichartige Bereiche, diese stellen den automatisierten Teil des Lagers dar. Der automatisierte Teil ist in Form eines automatischen Hochregallagers realisiert, die Besonderheit liegt im Regalbediengerät. Dies funktioniert nicht wie im klassischen Hochregallager, sondern teilt sich in zwei einzelne Module auf. Eines davon fährt horizontal, nimmt die Kiste am Übergabepunkt auf, fährt sie in das Hochregallager und stellt sie dort auf einen Übergabeplatz. Von diesem Übergabeplatz nimmt das vertikal fahrende Modul die Kiste auf und stellt sie auf ihren vorgesehenen Lagerplatz.

Der grün markierte Bereich stellt den *manuellen Teil* des Lagers dar, dort sind mehrere Etagen mit einer Vielzahl an Regalen bestückt. Jedes Regal ist mit Hilfe eines Kommissionierwagens

zugänglich und der Mitarbeiter kann die gewünschten Produkte in eine Transportkiste kommissionieren.

Beide Lagerbereiche müssen zusammengeführt werden, dafür existiert der *Picking Sorter* (gelbe Einfärbung). Dort werden die Aufträge konsolidiert, gepackt und an den Versandbereich übergeben. Die Prozesse werden in den beiden nachfolgenden Abschnitten genauer erläutert. Aus diesen Prozessen können Restriktionen entstehen, welche am Ende des Abschnittes genauer erläutert werden.

4.1.1.1 Materialfluss

Aus dem im oberen Abschnitt beschriebenen Aufbau eines teilautomatisierten Logistikzentrums resultieren Materialflüsse, welche in diesem Abschnitt näher beschrieben werden sollen. Zur Beschreibung der jeweiligen Abläufe wird zur besseren Übersichtlichkeit das Prozessmodell nach [Kuh95] verwendet. Die einzelnen Elemente werden in dieser Arbeit nicht erläutert, dafür wird auf [Kuh95] verwiesen. In Abbildung 14 ist der Materialfluss zu erkennen. Es kann zwischen vier großen Bereichen unterschieden werden, im Vergleich zum vorherigen Abschnitt waren es jedoch sechs. Die automatisierte Teil, das Pick-Modul und das Nachfülllager wurden zu dem Bereich Lagerung zusammengefasst. Der in der Abbildung 14 obere Prozess entspricht den vier Bereichen im Lager, die Prozesse in diesen Bereichen gliedern sich jeweils und die darunter stehenden Prozesskettenelemente auf. Die Farben entsprechen denen, die auch in der Abbildung 13 genutzt wurden. Im Nachfolgenden werden ausgehend von den vier oberen Prozesskettenelementen, alle unteren Prozesskettenelemente beschrieben. Der Materiafluss ist vereinfacht dargestellt, es wird sich nur auf die Bereiche die in der Problemstellung von Bedeutung sind bezogen.

Der *Materiafluss* beginnt mit dem Wareneingangsbereich, dort werden die Produkte aus dem LKW entladen. Nach der Entladung werden die Produkte zur Warenvereinnahmung transportiert. Bei der Warenvereinnahmung wird jedes Produkt mit einem internen Barcode zur eindeutigen Identifikation versehen. Sofern das Produkt noch nie im Lager vorhanden war, wird eine „First-Time-SKU“ durchgeführt. Dabei werden alle Abmessungen des Produktes, das Gewicht, ein Bild, die Kategorie und sonstige produktabhängige Daten in das Datenbanksystem eingespielt. Nach der erfolgreichen Warenvereinnahmung werden die Produkte über den Wareeingangssorter den verschiedenen Bereichen zugeteilt. Bei dieser Zuordnung wird vom WMS entschieden, in welchen Lagerbereich das jeweilige Produkt mit welche Menge eingelagert wird. Daher muss bereits in dieser Situation die bestehende Entscheidung geändert werden, um eine Optimierung zu erreichen. Neben dem beschriebenen Prozess für Produkte, die über die Fördertechnik transportiert werden können, existiert ein weiterer Prozess für übergroße Produkte. Diese werden in einen manuellen Bereich gebracht, dort vereinnahmt und zum Offline Picking mit integriertem Nachfülllager gebracht. Ebenfalls wird das Produkt dort manuell verpackt und zum Versandbereich transportiert und direkt auf den LKW geladen. Dieser Prozess spielt bei der Betrachtung der Einlagerung keine weitere Rolle.

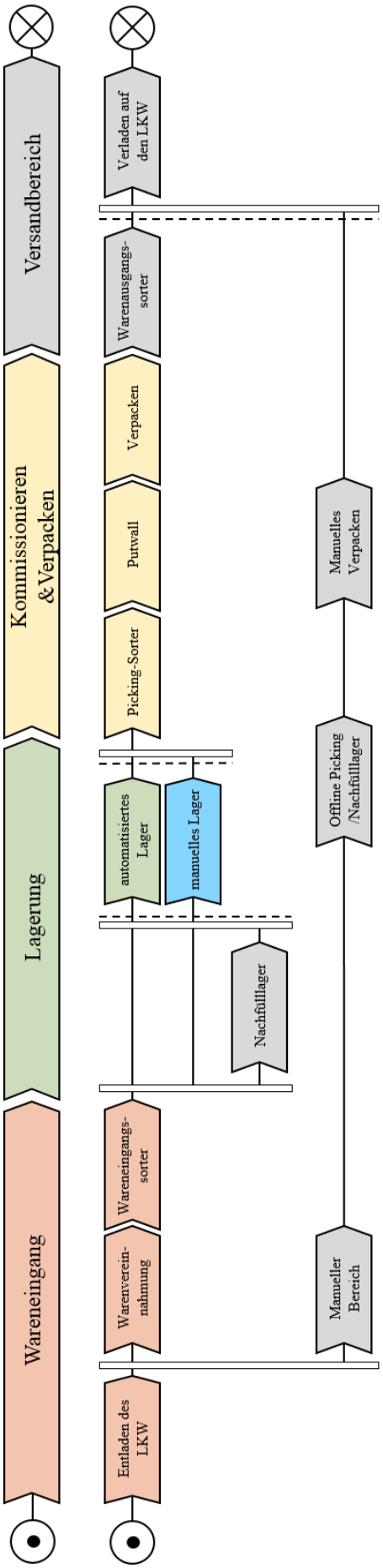


Abbildung 14: Materialfluss in einem teilautomatisierten Logistikzentrum

Die Produkte müssen nach der Aufteilung in die beiden Bereiche wieder zusammengeführt werden, dies geschieht im Bereich *Kommissionieren&Verpacken*. Dafür werden sie im manuellen Teil in Kisten kommissioniert und diese werden zum Picking Sorter über die Fördertechnik transportiert. Dabei sind die Kisten des manuellen Teils die gleichen Kissten, wie die aus dem automatisierten Teil. Die Kisten aus dem automatisierten Teil werden automatisiert ausgelagert und ebenfalls zum Picking Sorter automatisiert gefördert. Die Produkte werden an Ware-zu-Person Arbeitsplätzen per Hand auf einen Sorter gelegt, welcher die Produkte in einen vorbestimmten Bereich ausschleust. Nach der Ausschleusung, werden die Produkte zu den einzelnen Aufträgen in einer „Putwall“ zusammengefasst. Diese „Putwall“ ist ein Regal mit zahlreichen Fächern, wobei jedes Fach einen anderen Auftrag darstellt. Sofern der Auftrag komplett ist, wird er verpackt und dem Versandbereich übergeben.

Der Transport zum *Versandbereich* erfolgt voll automatisiert, dort werden die Produkte vorsortiert und in die entsprechenden LKW eingeladen. In diesem Bereich treffen sich das Offline Picking und der betrachtete Prozess wieder. Der Versandbereich ist der letzte Bereich, welche durchlaufen wird und das Produkt verlässt das Logistikzentrum.

Die Bereiche des Wareneingangs und der Lagerung sind für das vorliegende Problem interessant, da in diesem festgelegt wird, wie das Produkt eingelagert wird. Zu einem besseren Verständnis über die notwendigen Informationen in dem Logistikzentrum, wird im Folgenden der Informationsfluss vorgestellt.

4.1.1.2 Informationsfluss

Der Informationsfluss wird ausgehend vom Materiafluss mit den dazugehörigen Informationen beschrieben, welche die Auslöser für die jeweiligen Prozesse sind. Grundsätzlich kann zwischen zwei unterschiedlichen Startpunkten unterschieden werden, dies sind die Bestellung des Kunden und die Bestellung des Logistikzentrums beim Lieferanten. Das Prozesskettenmodell nach [Kuh95] in Abbildung 15 verdeutlicht die Problematik und zeigt den Ort der Zusammenführung dieser beiden Startpunkte. Der Aufbau gleicht dem aus Abbildung 14, somit beschreiben die vier oberen Prozesskettenelemente die Bereiche im Logistikzentrum, dass das jeweilige Produkt durchlaufen muss. Darunter werden die Informationsflüsse in den einzelnen Bereichen abgebildet. Der Informationsfluss ist auf Grund der Kapazität der Arbeit vereinfacht, es werden nur die Wesentlichen Schritte zur Lösung des Problems aufgezeigt.

Im Bereich des *Wareneingangs* werden drei Kerninformationen verarbeitet, zu Erst muss das Produkt im System angelegt werden, dafür müssen alle notwendigen Produktdaten eingegeben werden. Diese Information wird genutzt, um den Lagerbereich auf Basis der vorliegenden Daten zu bestimmen. Daraufhin wird ein Label (Barcode) zur eindeutigen Identifikation erzeugt und dieser auf das Produkt geklebt. In diesem Punkt kommen verschiedene Restriktionen zum Tragen.

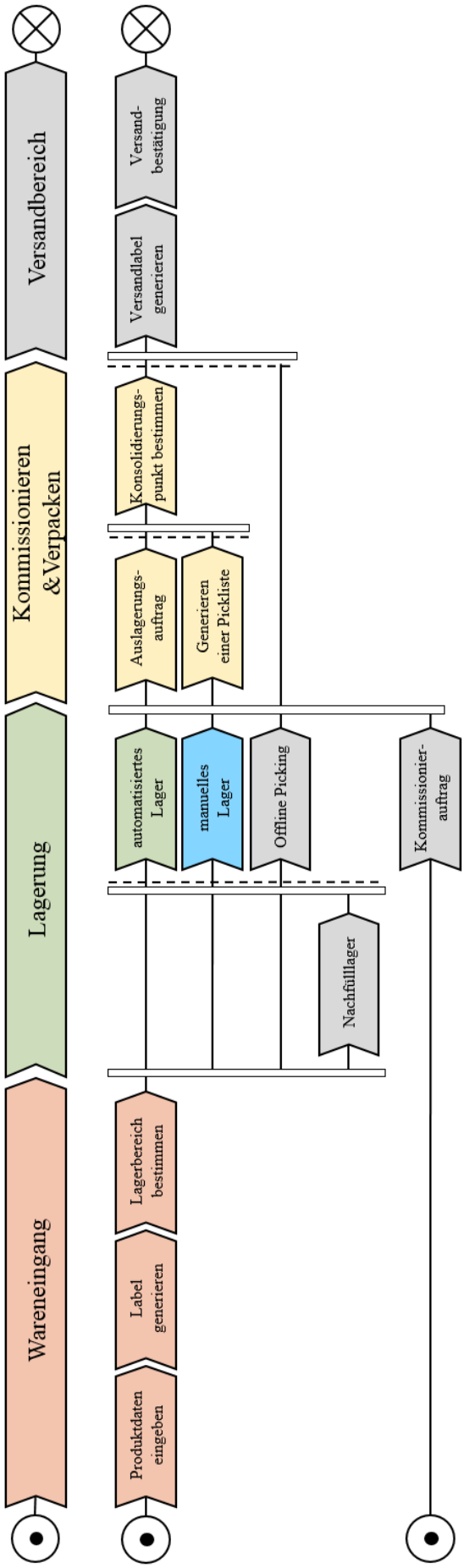


Abbildung 15: Informationsfluss in einem teilautomatisierten Logistikzentrum

Nach der Bestimmung des Lagerbereiches, werden die Produkte im Bereich *Lagerung* eingelagert. Dafür wird die Information abgespeichert, wo das Produkt gelagert wird. Entweder in den automatisierten Teil, im manuellen Teil, im Nachfülllager oder im Offline Picking. Sobald ein kritischer Lagerbestand in den Bereichen automatisiertes Lager, manuellen Teil oder Offline Picking erreicht ist, wird aus dem Nachfülllager in den jeweiligen Bereich umgelagert. Dabei wird ebenfalls wieder überprüft, ob es einen Bereichswechsel erfolgen soll.

Der bis hierhin beschriebene Prozess wird durch die Bestellung des Logistikzentrums beim Lieferanten ausgelöst. Nun wird der Prozess durch die Bestellung des Kunden ergänzt. Um ein erfolgreiche Auslagerung aus den einzelnen Bereichen zu gewährleisten, muss im ersten Schritt ein Kommissionierauftrag gebildet werden.

Auf Basis dieses Kommissionierauftrages werden die Prozesse im Bereich *Kommissionieren & Verpacken* gesteuert. Die Produkte aus dem automatisierten und dem manuellen Teil müssen konsolidiert werden, dafür wird für den manuellen Teil eine Pickliste für mehrere Aufträge erstellt. Sofern diese beendet wurde, werden die Produkte aus dem automatisierten Teil per Auslagerungsauftrag ausgelagert. Sind alle Produkte vorhanden, kann der Auftrag konsolidiert werden und verpackt werden und das Versandlabel generiert werden. Die Produkte des Offline Picking werden nicht konsolidiert sondern in ihrer eigenen Verpackung verschickt, daher wird für sie ebenfalls als nächstes nach der Auslagerung ein Versandlabel generiert.

Das Generieren des Versandlabels gehört *Versandbereich*, nachdem die Pakete mit dem Versandlabel versehen sind, wird dies einmalig beim Verladen in den LKW gescannt und somit eine Versandbestätigung generiert und an den Kunden versendet. Damit verlässt das Produkt das Logistikzentrum und der Informationsfluss ist beendet.

4.1.1.3 Restriktionen aus den Prozessen

Aus dem Layout, dem Materialfluss und dem Informationsfluss des teilautomatisierten Logistikzentrums resultieren verschiedene Restriktionen für die Lagerung der Produkte. Diese Restriktionen werden erläutert um sie in der Vorverarbeitung der DM-Verfahren zu berücksichtigen. Die Restriktionen lassen sich in folgende Gruppen einteilen:

- Abmessungen
- Beschaffenheit des Produktes
- Gewicht

Die erste Restriktion sind die *Abmessungen* des jeweiligen Produktes. Mit Hilfe der Abmessungen wird bestimmt, ob das Produkt über die automatisierte Fördertechnik transportiert werden kann oder nicht. Aufgrund dieser Restriktion, wird direkt im Wareneingang entschieden, ob das Produkt in den Bereich des Offline Pickings gelangt oder dem beschriebenen Prozess zugeführt wird. Die Entscheidung beruht auf den Abmaßen für eine Förderkiste. Diese Kiste hat definierte Abmessungen und jedes Produkt das größer ist, kann nicht in diese Kiste eingelagert werden. Da in dieser Arbeit die Produkte entweder in den manuellen oder den automatisierten Teil eingelagert werden sollen, muss das Produkt immer in die Kiste passen. Alle anderen Produkte werden nicht weiter betrachtet, da sie aufgrund ihrer Abmessungen in keiner der beiden Bereiche transportiert werden können.

Ebenfalls ist die *Beschaffenheit* der Produkte eine eingrenzende Restriktion. Dafür existieren verschiedene Lagermöglichkeiten für die einzelnen Produkte. Es kann zwischen sechs Arten unterschieden werden. Die Produkte mit einem hohen Wert dürfen nur in speziell überwachten Bereichen im manuellen Teil gelagert werden. In den automatisierten Teil hat das keine Auswirkungen, da die Regale nicht vom Personal erreichbar sind. Sofern die Produkte chemisch sind, dürfen diese im manuellen Teil, sowie auch in dem automatisierten Teil nur auf der untersten Ebene gelagert werden. Zu den chemischen Produkten zählen unter anderem Haarwaschmittel oder Reinigungsmittel. Die chemischen Produkte können entflammbar sein, sofern sie diese Eigenschaft erfüllen, müssen sie im Gefahrgutlager vorgehalten werden. Sofern die Produkte flüssig sind, dürfen diese in dem automatisierten Teil nur auf der untersten Ebene gelagert werden, für den manuellen Teil hat das keine weiteren Auswirkungen. Ein besonderen Lagerbereich ist für ölige Produkte (z.B.: Motoröl) festgelegt. Diese Produkte werden in besonderen Regalen untergebracht. Die letzte Gruppe stellen alle Produkte dar, die keinen besonderen Lagerbereich benötigen. Diese können an jeder Stelle im manuellen Teil oder in den automatisierten Teil eingelagert werden. Die Produkte für das Gefahrgutlager können bereits ausgeschlossen werden, denn für diese kann keine Entscheidung über die Einlagerung optimiert werden. Ähnlich verhält es sich mit den öligen Produkten, diese werden ebenfalls ausgeschlossen und nicht weiter in der Optimierung der Einlagerung betrachtet.

Den letzten Bereich stellt das Gewicht dar, in den automatisierten Teil können nur Kisten bis zu einem vorher definierten Gesamtgewicht eingelagert werden. Daher können die Transportkisten teilweise nicht komplett gefüllt werden.

Die letzte Restriktion stellt die Kapazitätsgrenze des automatisierten Teils dar, denn es kann nur eine bestimme Anzahl an Kisten in der Stunde ausgelagert werden. Dies muss bei der Entwicklung der beiden Vorgehensmodelle berücksichtigt werden, denn sonst kann der notwendige Gesamtdurchsatz nicht erreicht werden.

Aus diesen Restriktionen werden im Folgenden Anforderungen an die Attribute definiert, welche in den Vorverarbeitungsschritten zu berücksichtigen sind. Damit nur die Datensätze betrachtet werden, die mit der Einlagerung in den manuellen und den automatisierten Teil in Verbindung stehen. Nachfolgend wird die Datenbankstruktur des teilautomatisierten Logistikzentrums erläutert.

4.1.2 Datenbankstruktur des Warehouse Management Systems

Die zu analysierenden Daten stammen aus einem WMS, welches eine feste Struktur besitzt. Diese Struktur wird in Form eines Entity-Relationship-Modells (ER-Modell) im Folgenden erläutert. Dabei ist zu berücksichtigen, dass nicht alle vorhandenen Tabellen im WMS aufgeführt sind, sondern lediglich die notwendigen für diese Arbeit. Im Modell sind fünf verschiedene Entitäten zu erkennen, welche in diesem Fall mit Tabellen gleichzusetzen sind. Eine reale Abbildung des gesamten WMS wäre für diese Arbeit nicht zuführend, da nicht mehr Daten zur Lösung der Problemstellung benötigt werden. Im Informationsfluss wurde bereits zwischen zwei Startpunkten unterschieden, diese finden sich in der Abbildung 16 jeweils auf der linken und rechten Seite wieder. Die Notation der einzelnen Entitäten wurde von den Beispieldaten übernommen. Die linke Seite zeigt den Kundenauftrag und die rechte Seite den Auftrag des Logistikzentrums an

den Lieferanten. Zur Bearbeitung des Kundenauftrages werden zwei Tabellen benötigt, die CustomerOrder und die CustomerOrderLine.

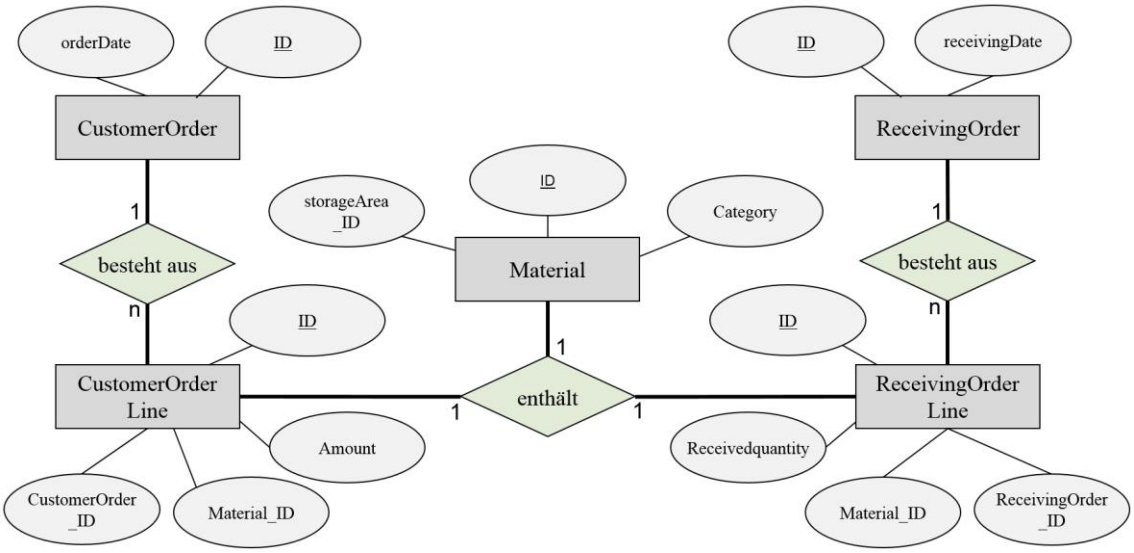


Abbildung 16: Entity-Relationship-Modell eines teilautomatisierten Logistikzentrums

In der Tabelle CustomerOrder werden alle Informationen über den Kunden und seinen Auftrag gespeichert. Diese Tabelle besitzt eine eindeutige ID und Informationen, wie das erwartete Versanddatum, oder das Datum des Eingangs der Bestellung (orderDate). Die CustomerOrderLine wiederrum macht Angaben über die bestellten Produkte, somit wird für jedes bestellte Produkt des Kunden ein Datensatz in der CustomerOrderLine angelegt. Diese Tabelle beinhaltet Informationen über die Menge (Amount), sowie eine ID zur Materialtabelle und zur CustomerOrder Tabelle.

Ähnlich verhält es sich auf der Seite des Auftrages des Logistikzentrums für die Nachbestellung und Neubestellung von Waren. Dieser befindet sich auf der rechten Seite und beinhaltet die Tabellen ReceivingOrder und ReceivingOrderLine. Die ReceivingOrder gibt die Bestellung bei dem Lieferanten an und beinhaltet Informationen über das Lieferdatum, den Anlieferungstyp und eine eindeutige ID. Die Tabelle verweist auf die ReceivingOrderLine, dort sind die Informationen zu den möglichen verschiedenen Produkten eines Auftrages abgespeichert. Dies beinhaltet unter anderem die Menge, einen Verweis auf die Materialtabelle und eine eindeutige ID.

Beide Auftragstypen verweisen auf die Materialtabelle, in dieser Tabelle stehen alle Informationen, welche über das Produkt bekannt sind. Neben der Kategorie (Category) des Produktes, welche das Produkt in verschiedene Produktkategorien einordnet, ist dort der aktuelle Lagerort vorzufinden (storageArea_ID). Ebenfalls existiert eine eindeutige ID in der Materialtabelle, welche zugleich der Primärschlüssel ist. Die beschriebenen Tabellen beinhalten weit mehr Attribute als im ER-Modell aufgeführt, die Dimension dieser Daten werden im Folgenden Abschnitt vorgestellt und näher erläutert.

4.1.3 Struktur und Dimensionen der Daten

Der Aufbau des WMS wurde bereits erläutert, die Dimension und Struktur der dahinter stehen Daten jedoch noch nicht. Zur Struktur werden die jeweiligen Metadaten der Tabellen und zur Dimensionen die Anzahl der jeweiligen Attribute und Zeilen der Tabellen erläutert. Durch den bereits beschriebenen Informations-und Materialfluss entsteht eine Vielzahl an Daten, welche in

zwei unterschiedlichen Tabellen vorliegen. In Bezug auf das ER-Modell sind die Daten des Kundenauftrages und der Materialtabelle jeweils in einer Tabelle zusammengefasst. Die Daten der Bestellung des Logistikzentrums des Lieferanten sind ebenfalls mit der Materialtabelle zu einer Tabelle zusammengelegt. Eine komplette Liste der Attribute befindet sich im Anhang A2 und A3. An dieser Stellen werden nur Anforderungen an die Attribute erläutert, welche zur Entwicklung des Vorgehensmodells unabdingbar sind. Die Tabelle 6 zeigt eine Übersicht über die Anzahl der Attribute in den jeweiligen Tabellen der Beispieldaten.

Tabelle 6: Anzahl der Attribute je Tabelle

Name der Tabelle	Anzahl Attribute
Material	83
RecievingOrder	21
RecievingOrderLine	24
CustomerOrder	48
CustomerOrderLine	22
Summe	**198**

Dabei bleibt festzuhalten, dass allein in den fünf notwendigen Tabellen 198 Attribute existieren. Insbesondere für die produktabhängigen Parameter existiert eine breite Auswahl an Attributen, welcher in der Tabelle Material aufgeführt sind. Neben den Attributen müssen ebenfalls die Zeilen betrachtet werden. Das System gibt jedoch nur zwei Tabellen aus, die Tabelle der Kundenbestellungen (Anhang A3), welche alle Aufträge der Kunden beinhaltet. In der Tabelle der Lieferantenbestellungen (Anhang A2) werden alle Bestellungen des Logistikzentrums bei den Lieferanten abgebildet. Im Anhang A3 wird die Materialtabelle nicht gesondert erläutert, die Attribute für die Materialtabelle sind aus dem Anhang A2 zu entnehmen. In der Tabelle 7 ist die Anzahl der Datensätze der einzelnen Tabellen an den Beispieldaten aufgezeigt

Tabelle 7: Anzahl der Datensätze je Tabelle

Name der Tabelle	Anzahl Datensätze	Anzahl Attribute
Kundenbestellungen	7.079.265	153
Lieferantenbestellungen	468.447	134

Die Anzahl der Datensätze in der Tabelle der Kundenbestellungen ist hoch, daraus resultieren Probleme mit der Rechnerkapazität, wie in der Datenvorverarbeitung beschrieben. Die hohe Anzahl der Attribute stellt ebenfalls eine Herausforderung dar, da die Tabelle der Kundenbestellungen dadurch über eine Milliarde Zellen hat. Bei der Tabelle der Lieferantenbestellungen ist die Anzahl der Attribute ebenfalls hoch, jedoch die Anzahl Datensätze wesentlich geringer. Daher stellt dies keine Herausforderung in der Handhabbarkeit dar.

Eine Reduzierung der Daten ist notwendig, weil sonst keine Möglichkeit besteht sie mit vorhandener Rechnerkapazitäten durch DM-Verfahren bearbeiten zu lassen. Um diese Reduzierung vornehmen zu können, müssen Anforderungen an die Daten definiert werden. Diese Anforderungen werden in den Schritt der Vorverarbeitung übernommen und dort mit den vorhandenen Daten abgeglichen. Auf Basis der Anforderungen werden die geeigneten Attribute zur Lösung des Problems festgelegt. Folgende Anforderungen können definiert werden:

- Lagerbereich
- Bestelldatum
- Bestellmenge
- Eindeutige Identifizierung eines Produktes
- Eigenschaften der Produkte

Die Anforderung des *Lagerbereiches* gibt an, ob das Produkt beispielsweise in den automatisierten oder manuellen Teil des Lagers eingelagert werden soll. Diese Information ist notwendig, da mit Hilfe dieser der neue Lagerbereich bestimmt werden kann. Das *Bestelldatum* und die *Bestellmenge* sind in der Tabelle der Kundenbestellungen vorzufinden, diese beiden Anforderungen werden benötigt um die genaue Anzahl der bestellten Produkte und den zeitlichen Verlauf der Bestellungen zu betrachten. Um die genaue Anzahl der bestellten Produkte herauszufinden, müssen die einzelnen *Produkte jeweils eindeutig identifizierbar* sein. Dadurch wird verhindert, dass Produkte doppelt betrachtet werden. Durch die doppelte Betrachtung kann die Anzahl der bestellten Produkte negativ beeinflusst werden. Die letzte Anforderung wird von den Eigenschaften des Produktes bestimmt, diese sind besonders wichtig, da in dieser Anforderung mindestens die Attribute vertreten sein müssen, welche benötigt werden um die Restriktionen in den Daten abzubilden. Aus den Restriktionen geht hervor, dass die Abmessungen und das Gewicht gegeben sein müssen und ebenfalls die Beschaffenheit des Produktes. Diese möglichen Arten der Beschaffenheit lassen sich in Abschnitt 4.1.1.3 noch einmal genauer nachlesen. Diese Anforderungen werden benötigt um die Vorverarbeitungsphase der beiden Handlungsalternativen zu gestalten.

4.2 Vorhersage für den Lagerbereich

Die Vorhersage für den Lagerbereich ist die erste Handlungsalternative für die ein Vorgehensmodell vorgestellt und entwickelt wird. Auf Basis von Abbildung 11 wird dieses Vorgehen in Abbildung 17 noch einmal detailliert dargestellt und definiert gleichzeitig den Ablauf für diesen Abschnitt.

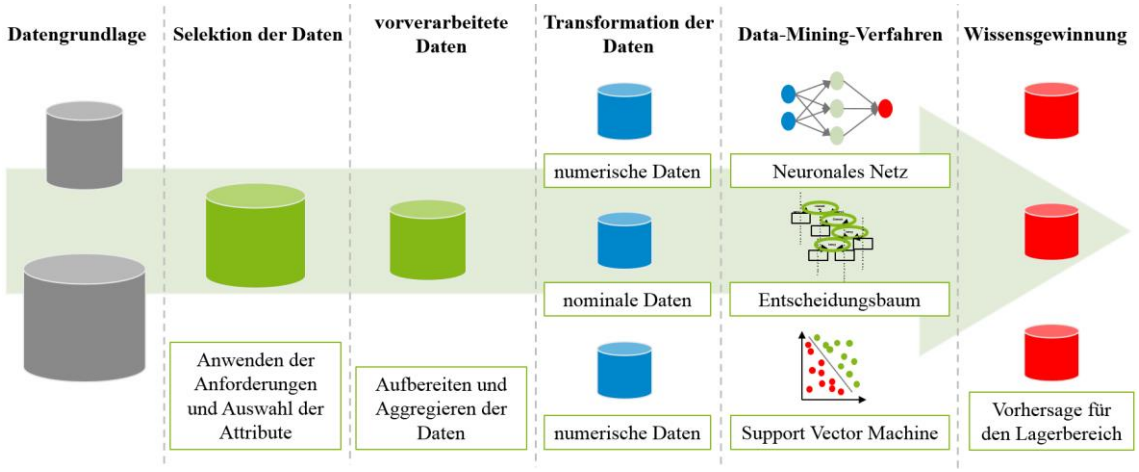


Abbildung 17: Vorgehensweise der Handlungsalternative Eins

Dabei steht der Ablauf des Vorgehensmodells von [FPS96] im Vordergrund, da nach diesem Schema ein Vorgehensmodell für die Problemstellung entwickelt werden soll. Die Daten liegen in zwei unterschiedlich großen Tabellen vor, dies stellt die Datengrundlage dar. Auf Basis dieser

Datengrundlage werden mit Hilfe der Restriktionen und den daraus resultierenden Anforderungen eine horizontale, sowie auch eine vertikale Reduktion der Dimensionen vorgenommen.

Nach der Auswahl der benötigten Datensätze und Attributen werden diese aufbereitet, indem fehlende Werte ergänzt und neue Attribute aggregiert werden. Die verschiedenen DM-Verfahren benötigen unterschiedlich transformierte Daten, daher werden zwei Datensätze mit numerischen Werten erzeugt und ein Datensatz mit nominalen Werten. Diese Daten werden daraufhin in die einzelnen DM-Verfahren überführt. Dafür werden ein NN, ein Entscheidungsbaum und eine SVM trainiert und auf die Daten angewendet. Diese Modelle lassen sich durch unterschiedliche Kennzahlen vergleichen, auf Basis dieser Kennzahlen und Expertenwissen wird ein Verfahren ausgewählt und zur Vorhersage des Lagerplatzes genutzt.

Bei der Durchführung des Vorgehensmodells ist permanent zu berücksichtigen, dass das Modell im Bereich des DM mit Trainingsdaten angelernt werden muss. Daher ist es unabdingbar geeignete Trainingsdaten zu haben, um eine validierbare Lösung zu erlangen. Diese Trainingsdaten müssen den Anspruch nach einer korrekten Zuordnung der Produkte zum jeweiligen Lagerbereich erfüllen. Der Ursprung der Trainingsdaten ist abhängig von dem jeweiligen Logistikzentrum und der daraus resultierenden Problemstellung, sie können beispielhaft über Simulationen erzeugt werden. Daher kann es durchaus vorkommen, dass diese Handlungsalternative nicht durchführbar ist, sofern keine geeigneten Trainingsdaten vorliegen. Der Begriff der Trainingsdaten wird im Folgenden nochmals verwendet, jedoch mit einer anderen Bedeutung. Für eine bessere Unterscheidung werden die in diesem Abschnitt beschriebenen Trainingsdaten als global gekennzeichnet und die im späteren Verlauf genutzten Trainingsdaten als lokal. Im Folgenden werden die einzelnen Vorverarbeitungsschritte näher erläutert und mit den Beispieldaten veranschaulicht.

4.2.1 Vorverarbeitung der Daten

Die Vorverarbeitung der Daten für die Vorhersage für den Lagerbereich stellt sich als Herausforderung dar, weil eine hohe Dimension der Daten vorliegt. Insbesondere ist durch die Rechenkapazität die Anzahl der nutzbaren Datensätze begrenzt, deswegen muss zu Beginn der Datenvorverarbeitung ein geeigneter Weg gefunden werden die Daten so zu reduzieren, damit sie in einem DM-Programm nutzbar sind.

Die Vorverarbeitung wird in drei Schritten durchgeführt, wie in Abbildung 17 dargestellt. Mit den unbearbeiteten Daten wird im ersten Schritt eine Selektion vorgenommen. Mit diesen selektierten Daten werden Aggregationen und andere vorverarbeitende Maßnahmen auf die Daten angewendet um sie im letzten Schritt für die jeweiligen DM-Verfahren zu transformieren. Eine genaue Abarbeitung der einzelnen Schritte gestaltet sich als kompliziert, da das KDD ein iterativer Prozess ist. Deswegen können zu einem späteren Zeitpunkt ebenfalls Daten selektiert werden bzw. Daten zu Beginn schon transformiert werden. Das entwickelte Vorgehensmodell muss als iterativer Prozess verstanden werden. Die Selektion der Daten wird im folgenden Abschnitt behandelt.

4.2.1.1 Anwendung der Anforderungen an die Daten

Die Daten müssen vollständig in ein geeignetes System eingelesen werden, um die genaue Selektion der Daten durchführen zu können. In Abbildung 18 sollen die beiden grauen zylindrischen Formen die Tabellen symbolisieren. Die obere Tabelle enthält die Daten der Bestellungen des Logistikzentrums bei dem Lieferanten (Nachschubbestellungen) und in der unteren Tabelle sind die Bestellungen der Kunden beim Logistikzentrum symbolhaft dargestellt.

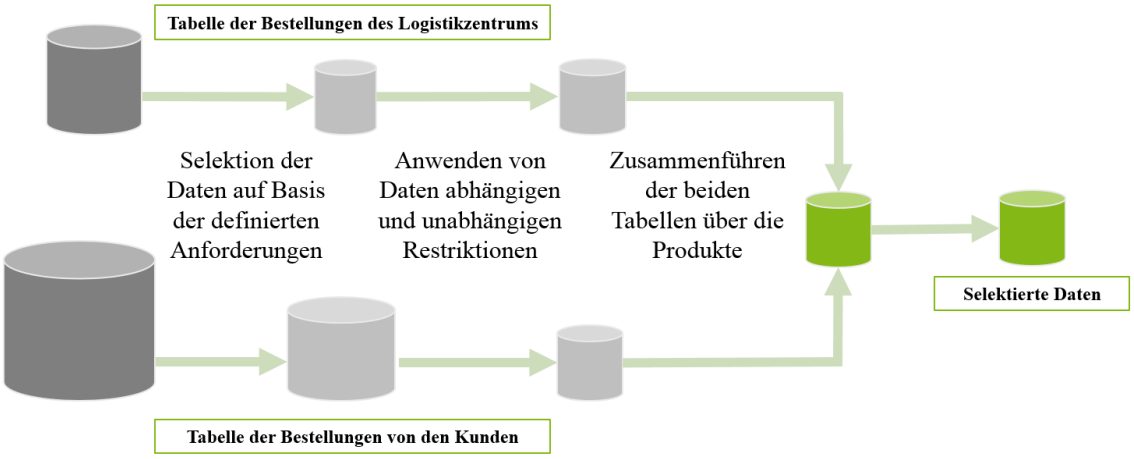


Abbildung 18: Selektion der Daten in Handlungsalternative Eins

Im ersten Schritt muss auf Basis der definierten Anforderungen eine *Selektion der Daten vorgenommen werden*, um überhaupt Berechnungen vornehmen zu können. Diese Selektion wird im ersten Schritt auf vertikaler Ebene durchgeführt, indem die Attribute betrachtet werden. Eine Selektion bedeutet ebenfalls eine Dimensionsreduktion, welche damit gleichzusetzen ist. In Anbetracht der Problemstellung können die Attribute herausgelöscht werden, welche die Folgenden Inhalte besitzen:

- Angaben über die letzte Änderung des Datensatzes
- Attribute, welche von übergeordneten IT-Systemen kommen, jedoch keine Funktion in der Tabelle haben
- Attribute, welche nicht direkt in Zusammenhang mit der Aufgabenstellung stehen
- Redundante Attribute, welche durch das Zusammenführen von Tabellen entstehen können (meistens an einem Index hinter dem Attributnamen zu erkennen)
- Attribute bei denen alle Werte fehlen

Von den beschriebenen Inhalten sind die meisten eindeutig. Die Anwendung der einzelnen Anforderungen setzt voraus, dass die Metadaten bekannt sind. Die Metadaten beschreiben die Attribute und geben Auskunft darüber, wie viele fehlende Werte oder welchen Wertebereich die jeweiligen Attribute besitzen. Es sind alle Inhalte eindeutig, bis auf die Anforderung danach, dass Attribute gelöscht werden, welche nicht direkt in Zusammenhang mit der Aufgabenstellung stehen. Dabei handelt es sich im produktspezifische Attribute und Attribute, welche eine Aussage über den Lagerbereich treffen. Dabei hängen die produktspezifischen Attribute mit den Anforderungen an die Daten aus Abschnitt 4.1.3 zusammen. Daher müssen alle Attribute, welche in irgendeiner Form die Restriktionen beschreiben in der jeweiligen Tabelle bestehen bleiben. Neben der vertikalen Selektion, werden die Daten ebenfalls horizontal selektiert. Dafür wird die Anforderung nach einer eindeutigen ID für jedes einzelne Produkt benötigt. Jedes Produkt wird nicht nur einmal an einen Kunden verkauft, sondern mehrmals an verschiedene Kunden. Somit ist in

der Auflistung der Tabelle der Kundenbestellungen ein Produkt häufiger vertreten. Deswegen werden alle doppelten Produkte aus der Tabelle herausgelöscht. Damit ist jede ID des Produktes in der Tabelle einzigartig, mit dem Reduzieren der Dimensionen in der Tabelle der Kundenbestellungen entsteht ein Informationsverlust. Dadurch lässt sich die Häufigkeit des verkauften Produktes nicht mehr ermitteln. Zur des Informationsverlustes wird im nachfolgenden Abschnitt eine Methode vorgestellt. Weitere horizontale Dimensionsreduktionen werden mit der Tabelle der Kundenbestellungen nicht durchgeführt. Neben der Tabelle der Kundenbestellungen besteht ebenfalls die Tabelle der Lieferantenbestellungen. Bei der Tabelle der Lieferantenbestellungen werden ebenfalls die doppelten Werte für die Produkte herausgelöscht, damit die ID für die Produkte einzigartig ist. Der daraus entstandene Informationsverlust ist nicht weiter zu betrachten, weil die eingehenden Produkte für die Bestimmung des Lagerbereiches irrelevant sind. Denn die Bestimmung des Lagerbereiches wird auf Basis der Häufigkeit der jeweiligen Auslagerungen der einzelnen Produkte bestimmt. Dabei ist zu berücksichtigen, dass nur zeitunabhängige Daten betrachtet werden und in die Tabelle überführt werden. Da jedoch eine Vorhersage getroffen werden muss, müssen noch zeitabhängige Daten der Tabelle zugeführt werden.

Nachdem eine Selektion stattgefunden hat, werden die Daten auf Basis der definierten Restriktionen weiter vorverarbeitet. Dabei besteht der Hauptunterschied, dass bei der Selektion gesamte Attribute oder gesamte Zeilen gelöscht werden, im Schritt des *Anwendens der Restriktionen* werden einzelne Werte aus den Daten herausgefiltert. Dabei muss bei der Tabelle der Lieferantenbestellungen betrachtet werden, ob alle aufgeführten Produkte bereits im Lager eingetroffen sind. Sollten die Produkte noch nicht im Lager eingetroffen sein, können sie ebenfalls aus der Tabelle gelöscht werden. In der Tabelle der Kundenbestellungen können unterschiedliche Typen an Bestellungen vorliegen, welche das DM-Ergebnis verfälschen. Deswegen werden nur Bestellungen betrachtet, die vom Kunden getätigt wurden und systemseitig bedingte Bestellungen werden gefiltert. Zu den systemseitig bedingten Bestellungen kann eine Lieferung von Waren aus einem Logistikzentrum in ein anderes Logistikzentrum gehören. Das Auftreten von fehlenden Werten am Ende des Selektionsvorganges benötigt weitere Filtermaßnahmen. Dazu muss ein Verständnis für die Prozesse und die Daten im speziellen Logistikzentrum vorliegen. Dabei können beispielsweise die Datensätze bei denen die Abmessungen fehlen ausgeschlossen werden. Aus den Prozessen resultierend darf kein Produkt eingelagert werden, bei dem die Abmessungen nicht vollständig eingegeben sind.

Die letzte Aufgabe besteht darin, die *Tabellen zusammenzuführen*. Im Lager liegen Produkte, welche noch nie verkauft wurden. Produkte die noch nie verkauft wurden, sind in der Tabelle der Lieferantenbestellungen zu finden, Produkte welche schon verkauft wurden in der Tabelle der Kundenbestellungen. Somit wird ein Abgleich zwischen den beiden Tabellen gebildet und daraus resultiert eine Tabelle mit allen Produkten, welche sich derzeit im Lager befinden. Diese zusammengeführte Tabelle stellt das Ergebnis des ersten Vorverarbeitungsschrittes dar.

4.2.1.2 Aggregation der Daten

Durch die vertikale aber insbesondere horizontale Dimensionsreduktion wurden Informationen gelöscht. Diese sollen im Rahmen der Aggregation den Daten zurückgeführt werden, ebenfalls werden neue Attribute generiert um die beschriebenen Restriktionen anzuwenden. In Abbildung 19 ist das Vorgehen für diesen Abschnitt dargestellt.

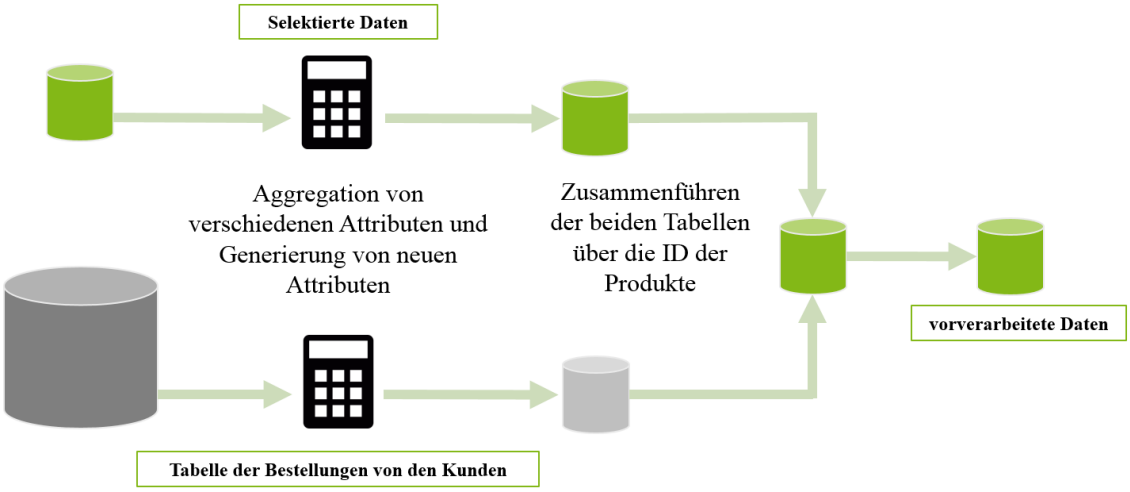


Abbildung 19: Aggregation der Daten in Handlungsalternative Eins

Das Vorgehen beginnt mit dem Bezug zum vorherigen Abschnitt, denn die bereits selektierten Daten werden benötigt um daraus ein neues *Attribut zu aggregieren*. Diese sind im oberen Bereich der Abbildung 19 zu erkennen, im unteren Bereich wird nochmals mit den gefilterten Ursprungsdaten begonnen. Dabei bedeutet gefiltert, dass die oben beschriebenen Restriktionen auf die Daten angewendet werden, jedoch noch nicht die vertikale oder horizontale Selektierung. Denn der generierte Informationsverlust soll mit der Aggregation von einem neuen Attribut abgefangen werden. Um dieses neue Attribut zu aggregieren, müssen jedoch die gefilterten Daten vorliegen, denn es sollen keine falschen Bestellungen enthalten sein. Mit der Aggregation wird das Ziel verfolgt, die Anzahl jedes einzeln verkauften Produkts über die Zeit darzustellen. Dadurch wird ebenfalls ein zeitabhängiges Attribut den selektierten Daten zugeführt. Bei der Aggregation werden alle Mengen der jeweiligen Produkte summiert, nach dem jeweiligen Produkt gruppiert und in eine Tabelle geschrieben. Damit existiert für jedes Produkt eine verkaufte Menge über einen Zeitraum. Auf Basis dieser Menge kann nun die Häufigkeit des einzelnen Produktes für den betrachteten Zeitraum errechnet werden. Da diese Tabelle als eindeutige ID das Produkt hat, kann sie mit den selektierten Daten zusammengeführt werden.

Mit den selektierten Daten wird ebenfalls ein *neues Attribut generiert*. Mit diesem Attribut soll das Volumen für jedes Produkt bestimmt werden. Das Volumen berechnet sich aus der Multiplikation der Höhe, Breite und Tiefe. Somit müssen in der Datenselektion die Abmessungen mit in die selektierten Daten übernommen werden. Dabei muss jedoch berücksichtigt werden, dass die Zahlenangaben unterschiedlichen Einheiten unterliegen können. Das ist abhängig von dem Standort des jeweiligen Logistikzentrums. Sofern eine andere Einheit vorliegt, als die Benötigte, muss diese noch umgerechnet werden, bevor sie zur Berechnung des Volumens genutzt werden kann. Es besteht die Möglichkeit, dass die Daten bereits ein Attribut mit dem Werten des Volumens haben, dann ist die Berechnung hinfällig.

Der letzte Schritt besteht darin, die beiden Tabellen über die ID der *Produkte zusammenzuführen*. Bei der Zusammenführung, wird es Produkte geben, welche einen fehlenden Wert im Bereich der Häufigkeit aufweisen. Das hängt damit zusammen, dass Produkte existieren die noch nie verkauft wurden (vgl. Abschnitt 4.2.1.1), für diese Produkte soll der Wert auf 0 gesetzt wer-

den, da die DM-Verfahren mit fehlenden Werten nicht umgehen können. Mit der Zusammenführung entstehen die vorverarbeiteten Daten, welche im nächsten Schritt noch transformiert werden müssen.

4.2.1.3 Transformation der Daten

Der letzte Vorverarbeitungsschritt besteht darin die Tabellen zu transformieren und letzte Restriktionen anzuwenden. In der Abbildung 20 ist das Vorgehen in diesem Schritt des KDD-Vorgehensmodells zu erkennen.

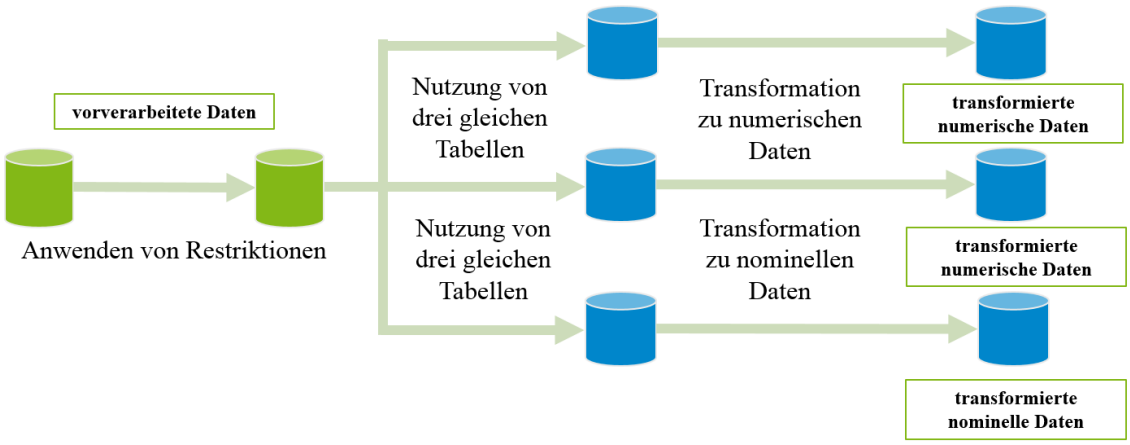


Abbildung 20: Transformation der Daten für Handlungsalternative Eins

Die Grundlage stellen die bereits vorverarbeiteten Daten dar. Der erste Schritt bei der Transformation der Daten besteht darin, letzte *Restriktionen anzuwenden*. Dazu gehören unter anderem die nicht betrachteten Lagerbereiche auszuschließen und die Restriktion nach den Abmessungen zu überprüfen. Dabei können Produkte ausgeschlossen werden, welche in ein Gefahrgutlager oder eine sonstige spezielle Lagerung benötigen. Ebenfalls müssen die Produkte in eine Kiste passen, um eine Einlagerung im automatisierten und manuellen Teil zu ermöglichen. Daher wird das Volumen mit dem maximal möglichen abgeglichen, sollte ein Produkt größer sein, darf es nicht in einer der beiden Lager transportiert werden. Sofern alle Restriktionen angewendet wurden, können die Daten aufgeteilt werden.

Die fertige Tabelle wird zweimal kopiert um sie den entsprechenden DM-Verfahren zuzuführen. Dabei werden drei DM-Verfahren betrachtet, zwei jedoch benötigen numerische Daten und eines benötigt nominelle Daten. Die Transformation von nominellen zu numerischen Daten erfordert eine genaue Kenntnis über die Daten, denn für jede Ausprägung eines Attributes wird ein neues Attribut angelegt. Sofern eine hohe Anzahl an Ausprägungen vorliegt, sollte überlegt werden das Attribut auszuschließen. Bei der Transformation von numerischen zu nominellen Daten ist dies nicht notwendig, da nicht für jede Ausprägung eines Attributes ein neues Attribut angelegt wird. Neben der Transformation der Attribute zu den jeweiligen Datentypen, darf das Attribut mit der Bestimmung des Lagerbereiches nur zwei Ausprägungen besitzen. Daher muss das Attribut mit der Angabe des Lagerbereiches zu einem Datentyp transformiert werden, welcher nur zwei Ausprägungen zulässt. In den meisten Fällen muss das Attribut den Datentyp binominal bekommen. Dies wird benötigt um einen Vergleich der unterschiedlich angewendeten DM-Verfahren durchführen zu können. Die transformierten Daten werden nach Beendigung des Schrittes des KDD-Vorgehensmodells den DM-Verfahren zugeführt.

4.2.2 Anwendung von Data-Mining-Verfahren

Das DM ist der Hauptschritt im KDD-Vorgehensmodell und benötigt vorverarbeitete Daten, welche auf das einzelne DM-Verfahren abgestimmt sind. In diesem Abschnitt wird erläutert, wie die drei verschiedenen DM-Verfahren auf die Daten angewendet werden können. In Abbildung 21 ist der Ablauf für diesen Abschnitt zu erkennen, wobei der Schritt der *Entwicklung von DM-Modellen* nochmals einen eigenen Ablauf besitzt.

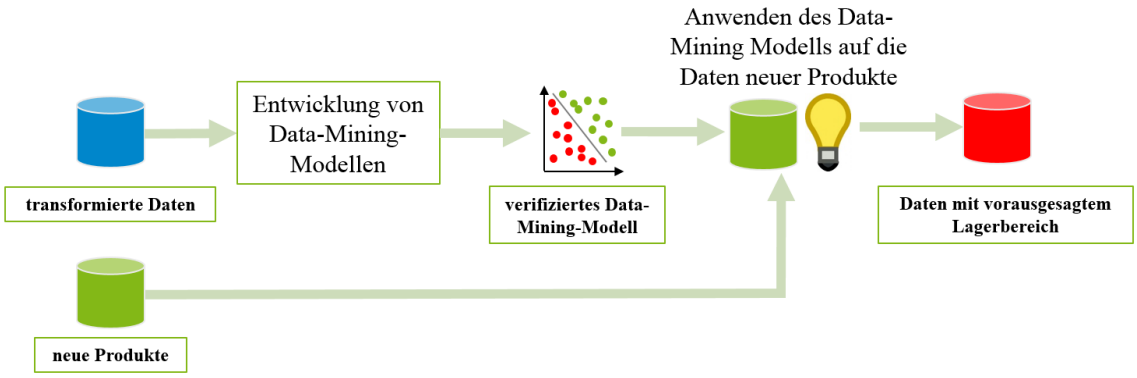


Abbildung 21: Anwendung von Data-Mining-Verfahren in Handlungsalternative Eins

Als Grundlage werden die transformierten Daten aus den Vorverarbeitungsschritten verwendet und den verschiedenen DM-Verfahren zugeführt. Die drei DM-Verfahren sollen jeweils validierte Modelle entwickeln, diese validierten Modelle werden auf die *Daten von neuen einzulagernden Produkten angewendet.*

Die Daten der neuen Produkte müssen die gleiche Struktur aufweisen, wie die transformierten Daten. Dabei unterscheiden sich die Daten darin, dass die neuen Produkte noch keine Angabe für den Lagerbereich haben. Auf Basis des erlernten Modells erfolgt die Zuordnung der Datensätze der neuen Produkte zu einem der beiden Lagerbereiche, entweder in den manuellen Teil oder den automatisierten Teil.

Diese Tabelle stellt das Ergebnis mit den *vorausgesagten Lagerbereichen* dar und kann in das WMS übernommen werden. Vor der Übernahme in das WMS muss jedoch eine Validierung der Ergebnisse durch eine Simulation durchgeführt werden. Die Ergebnisse der Simulation lassen sich wiederum in die Phase der Entwicklung des DM-Modells zurückführen und es kann eine Anpassung der Parameter erfolgen. Zur Erläuterung der möglichen Parameter wird im Folgenden der Ablauf von der Entwicklung eines DM-Modells beschriebenen.

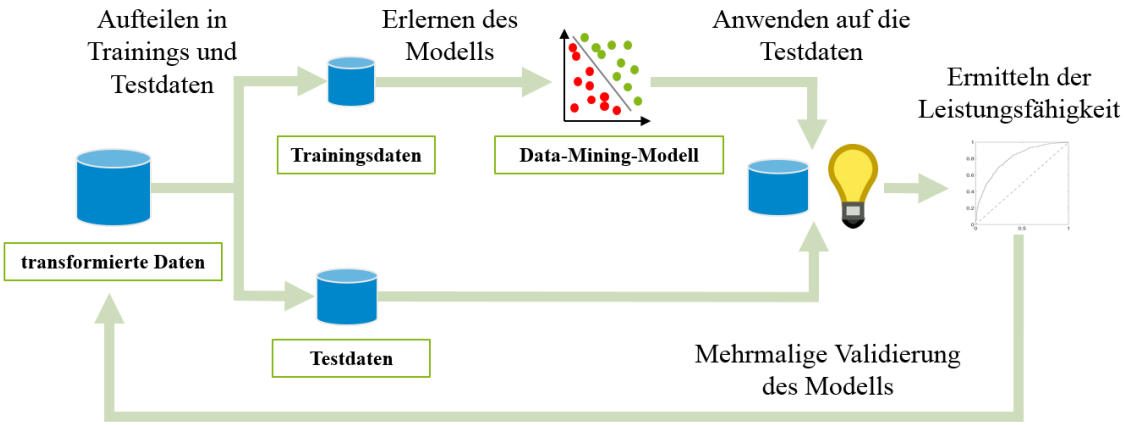


Abbildung 22: Ablauf der Data-Mining-Verfahren in Handlungsalternative Eins und Zwei

Der Ablauf der Entwicklung von einzelnen DM-Verfahren wird detaillierter beschrieben, weil es die Kernaufgabe des Vorgehensmodells darstellt. Der Beginn stellen ebenfalls, wie in Abbildung 21, die transformierten Daten dar, diese dienen als Basis zur Anwendung der DM-Verfahren mit dem Ziel verschiedene DM-Modelle zu erhalten. Dabei stellen die transformierten Daten im globalen Verständnis die Trainingsdaten für die Anwendung der DM-Verfahren dar. In diesen Daten sind die Produkte bereits den richtigen Lagerbereichen zugeordnet. Insgesamt werden drei verschiedene DM-Verfahren angewendet, bei allen herrscht das gleiche Vorgehen, daher wird eines beispielhaft beschrieben. Der einzige Unterschied besteht in den zugeführten Daten, diese können entweder nominell oder numerisch sein. Die numerischen Daten werden für die SVM und das NN verwendet und die nominellen Daten für den Entscheidungsbaum. Bevor die Daten den unterschiedlichen Verfahren zugeführt werden können, muss ein Zielattribut bestimmt werden, dies ist in diesem Fall der Lagerbereich. Eine weitere Besonderheit besteht darin, dass die auftretende Häufigkeit der Produkte als Gewicht in die Daten eingegeben wird. Denn den Werten im Attribut soll eine höhere Gewichtung zukommen, weil es der einzige zeitabhängige Faktor in der Tabelle ist.

Nach der Zuordnung der Rollen an die verschiedenen Attribute müssen die transformierten Daten *aufgeteilt werden in Trainings-und Testdaten*. Dabei besteht der Unterschied darin, dass die lokalen Trainingsdaten genutzt werden um das Modell zu erlernen und die Testdaten genutzt werden um das Modell zu validieren. Die Aufteilung erfolgt nach statistischen Methoden, daher ist sie zufällig und ändert sich bei jeder erneuten Aufteilung. Dabei müssen jedoch Trainings-und Testdaten die korrekte Zuordnung des Lagerbereiches zu den einzelnen Produkten enthalten.

In diesem Punkt setzt das eigentliche *DM* ein, denn es wird der Algorithmus des gewählten Verfahrens auf die lokalen Trainingsdaten angewendet. Eine weitere Herausforderung entsteht dadurch, dass der gewählte Algorithmus unterschiedliche vom Anwender zu bestimmende Parameter hat. Deswegen sollte der Anwender die Parameter automatisiert zu setzen, um somit den besten Wert herauszufinden. Dies erfordert eine hohe Rechendauer, daher sollte überlegt werden, dies nur mit einem Teil der Daten durchzuführen.

Nach der Entwicklung des Modells, wird dies auf die *Testdaten angewendet*. In diesem Schritt wird untersucht, wie genau das Modell die Testdaten zuordnet, daraus entsteht eine Möglichkeit zur Messung der Leistungsfähigkeit des entwickelten Modells.

Für die *Ermittlung der Leistungsfähigkeit* existieren verschiedene Methoden unter anderem der AUC. Der Wert des AUC wird gespeichert und es beginnt ein neuer Durchlauf der Entwicklung des DM-Modells.

Zur *mehrmaligen Verifizierung des DM-Modells* wird begonnen die transformierten Daten wieder in Trainings- und Testdaten aufzuteilen. Insgesamt wird eine vorbestimmte Anzahl an Durchläufen durchgeführt, diese Anzahl kann durch den Anwender festgelegt werden. Nach jedem Durchlauf lernt das Modell dazu und erreicht einen besseren AUC-Wert. Die unterschiedlichen Durchläufe dienen dazu, das Modell zu validieren und um eine bessere Leistungsfähigkeit im Vergleich zu nur einem Durchlauf zu erreichen. Sofern alle Verifizierungsdurchläufe beendet sind, wird ein Modell ausgegeben, welches auf die neuen Produktdaten anwendbar ist. Anhand dieses Modells werden nun die Lagerbereiche für die neuen Produkte bestimmt. Ebenfalls wird die Leistungsfähigkeit des Modells ausgegeben, diese wird genutzt um die Modelle untereinander vergleichen zu können.

Mit dem Vorhersagen des Lagerbereiches durch die unterschiedlichen DM-Verfahren sind alle notwendigen Schritte des Vorgehensmodells abgeschlossen. Dabei ist der letzte Schritt der Vorgehensweise nach [FPS96] bereits durch die Speicherung der Daten im WMS beendet. Das entwickelte Wissen wurde somit konserviert und ist für alle zugänglich.

Um eine Auswahl für das korrekte DM-Verfahren treffen zu können, wurden die Beispieldaten mit allen drei Verfahren durchgeführt. Die daraus resultierenden Ergebnisse finden sich im nachfolgenden Abschnitt wieder.

4.2.3 Auswahl eines Data-Mining-Verfahrens

Der Schritt der Auswahl eines DM-Verfahrens ist nur notwendig bei der Durchführung, wenn verschiedene DM-Verfahren durchgeführt werden. Daher ist dieser Schritt nicht in der Abbildung 17 zu erkennen. Sollte der Anwender sich für die Durchführung dieses Vorgehensmodells entscheiden dann kann er das in diesem Abschnitt ausgewählte DM-Verfahren verwenden. Dieses Verfahren hat bei der Lösung des Einlagerungsproblems die höchste Leistungsfähigkeit. Es muss jedoch darauf hingewiesen werden, dass die Leistungsfähigkeit abhängig von den Beispieldaten ist. Daher steht es dem Anwender frei ebenfalls die in diesem Abschnitt beschriebene Analyse durchzuführen, um das beste Ergebnis zu ermitteln.

Um eine Vergleichbarkeit der DM-Verfahren erreichen zu können, müssen folgende Kennwerte der Leistungsfähigkeit berechnet werden:

- Trefferwahrscheinlichkeit
- AUC
- ROC-Diagramm

Die Aussage der drei Kennwerte zur Bestimmung der Leistungsfähigkeit wurde bereits in Abschnitt 3.4 beschrieben. Im Folgenden werden jetzt mit Hilfe der Beispieldaten für die drei untersuchten Verfahren die Kennwerte aufgezeigt. Dafür werden zuerst alle Ergebnisse der drei Verfahren vorgestellt und anschließend mit einander verglichen. Es wird sich für das Verfahren entschieden, welches die besten Werte zur Messung der Leistungsfähigkeit erreicht.

Alle Abbildungen für den ROC sind gleich, daher werden sie nur einmal erläutert. Die jeweilig folgenden Abbildungen enthalten zwei Graphen, der rote Graph stellt den ROC dar und der blaue Graph den ROC (Thresholds). In dieser Arbeit ist nur der rote Graph interessant, er stellt

den in Abschnitt 3.4.3 vorgestellten Graphen dar. Der ROC (Thresholds) beschäftigt sich mit Schwellwerten, er wird nicht näher betrachtet. Auf der linken und rechten Seite vom roten Graphen, ist teilweise eine hellrote Fläche zu erkennen. Diese Fläche zeigt an, dass sich der Graph in diesem Korridor bewegen kann, abhängig davon ob ein optimistischer oder pessimistischer AUC-Wert vorliegt. Der AUC-Wert ist nicht direkt aus den Abbildungen ablesbar, er ist jedoch die Fläche unter der roten Kurve.

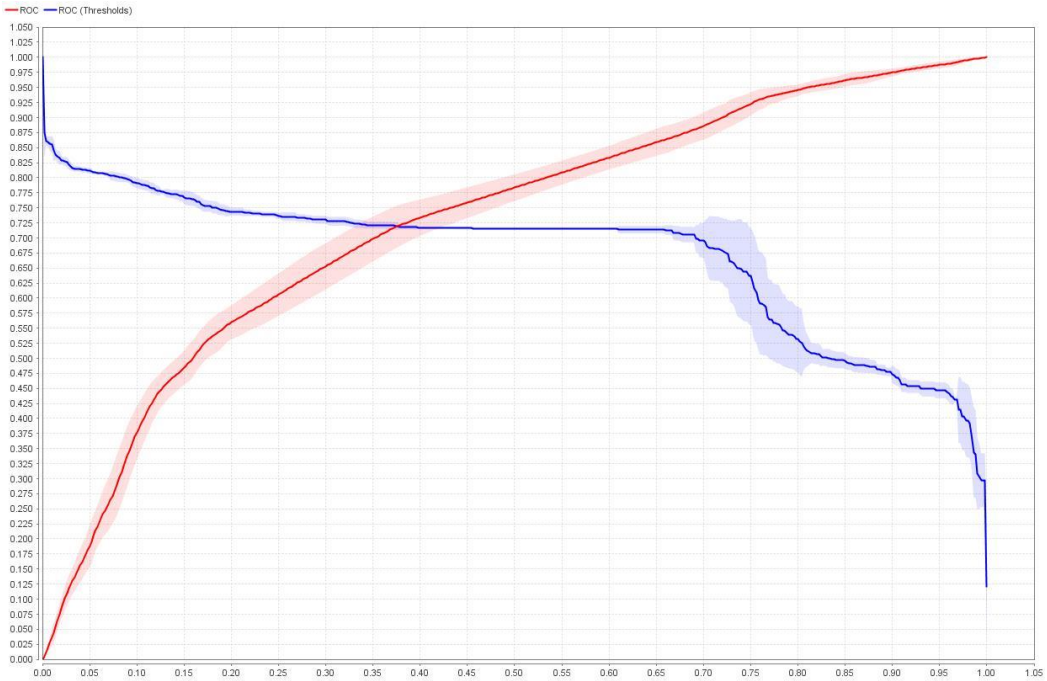


Abbildung 23: ROC des Entscheidungsbaumes

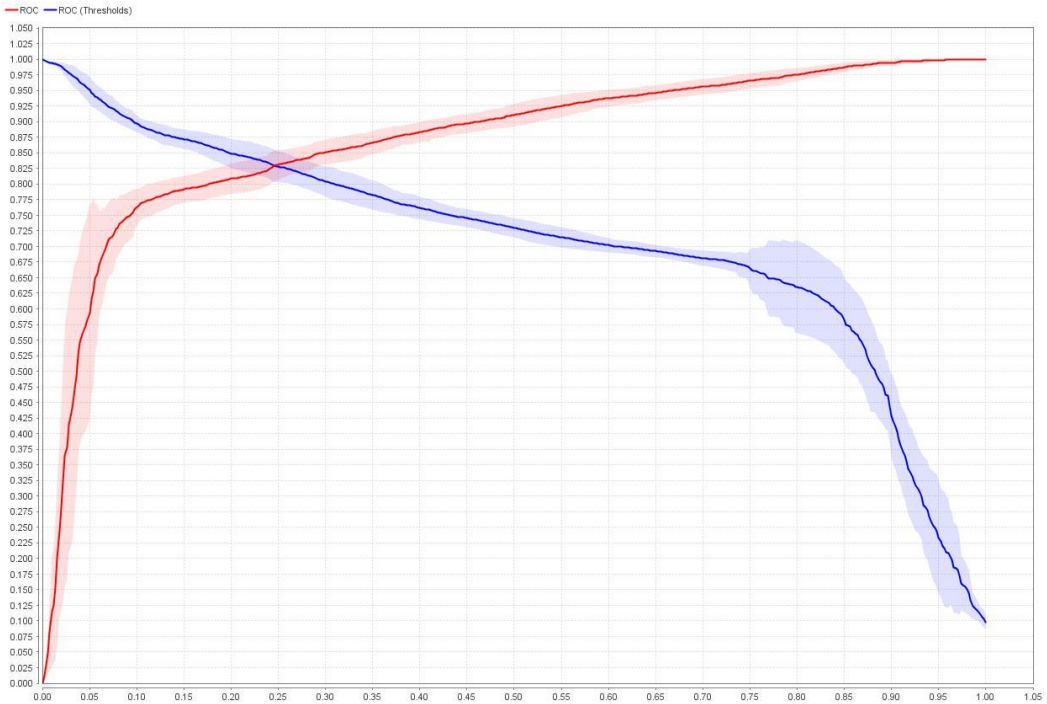


Abbildung 24: ROC des Neuronalen Netzes

In der Abbildung 23 ist der ROC des Entscheidungsbaumes zu erkennen, der AUC-Wert liegt bei 0,727 im Mittelfeld der möglichen Spanne des AUC. Ebenfalls ist der Korridor schmal gehalten in dem sich der rote Graph bewegt, deswegen sind die Ergebnisse der Klassifikation als eindeutig

zu bewerten. Der Graph bewegt sich konstant nach oben, ohne auffällige Schwankungen, was für eine erfolgreiche Klassifikation spricht. Neben dem AUC-Wert, hat das Modell des Entscheidungsbaumes eine Trefferwahrscheinlichkeit von 87,99% bei einer Abweichung von 1,13%. Die Trefferwahrscheinlichkeit ist hoch und die Abweichung gering, dies zeigt eine erfolgreiche Zuordnung der einzelnen Produkte zu den verschiedenen Lagerbereichen.

Das zweite zu untersuchende DM-Verfahren ist das NN für welches die Kennwerte der Leistungsfähigkeit vorgestellt werden. Die Abbildung 24 zeigt den ROC des NN, wobei direkt zu erkennen ist, dass die Kurve höher ist und der hellrote Korridor kleiner ist. Der AUC-Wert beträgt 0,872 und liegt damit höher, als der Wert vom Entscheidungsbaum. Durch diesen hohen Wert, ist die Klassifikation des NN besser gegenüber der Klassifikation des Entscheidungsbaumes. Bei der Trefferwahrscheinlichkeit verhält es sich ähnlich, diese liegt bei 91,34% mit einer möglichen Abweichung von 0,97%. Die Klassifikation mit dem NN ist nach der Aussage der Kennwerte zur Leistungsfähigkeit besser, als die des Entscheidungsbaumes.

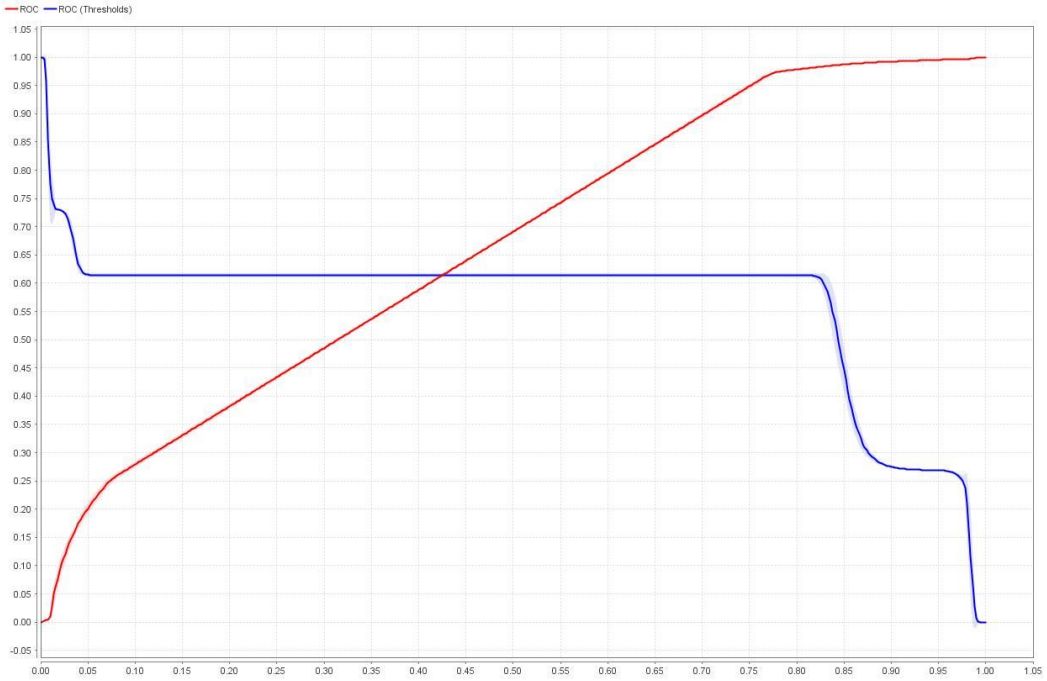


Abbildung 25: ROC der Support Vector Machine

Das letzte untersuchte Modell stellte die SVM dar, in Abbildung 25 ist der ROC zu erkennen. Dabei ist kein Korridor zu erkennen, welcher mögliche Abweichungen erlauben würde, generell ist die rote Linie jedoch wesentlich niedriger als bei den beiden bereits beschriebenen Verfahren. Daher ist der AUC auch wesentlich geringer, er liegt bei 0,663 bei einer Abweichung von 0,003. Ein niedriger AUC Wert, verursacht ebenfalls einen geringen Wert bei der Trefferwahrscheinlichkeit, dieser liegt bei 77,05% mit einer Abweichung von 0,12%. Die Abweichung stellt sich auch als geringer dar, als bei den oberen Verfahren. Somit stellt die SVM das schlechteste Verfahren im Vergleich von allen den drei Verfahren dar.

Tabelle 8: Vergleich der DM-Verfahren

DM-Verfahren	AUC	Trefferwahrscheinlichkeit
Entscheidungsbaum	0,727	87,99%
Neuronales Netz	0,872	91,34%
Support Vector Machine	0,663	77,05%

In der Tabelle 8 sind die Leistungskennwerte nochmals zum Vergleich aufgezeigt. Durch den Vergleich der beschriebenen Werte wird sich für das NN entschieden, weil es eindeutig die beste Trefferwahrscheinlichkeit und den höchsten AUC-Wert besitzt. Dies trifft auf die untersuchten Beispieldaten zu, sofern sich die Daten ändern sich die Entscheidung für eines der DM-Verfahren ändern.

4.2.4 Validierung des Vorgehensmodells

Für die Validierung des Vorgehensmodells existieren zwei verschiedene Möglichkeiten, Validierung durch Expertenwissen oder Validierung mit Hilfe einer Simulation. In dieser Arbeit sollen beide Verfahren vorgestellt werden und auf die Anwendbarkeit der Problematik geprüft werden.

Bei der Validierung durch Expertenwissen, wird ein mit den Beispieldaten und der Problemstellung vertrauter Experte benötigt. Dieser Experte nimmt händisch eine Zuordnung der Produkte zu den einzelnen Lagerbereichen vor. Dabei wird eine Zeit vorgegeben, in dieser Zeit muss der Experte so viele Beispiele wie möglich zuordnen. Daraufhin wird die Zuordnung mit denen des DM-Verfahrens verglichen und überprüft, welches Ausmaß eventuelle Abweichungen haben. Diese Art der Validierung ist bei einer Zuordnung in einem Logistikzentrum, bei denen über 100.000 Produkte auftreten keine effiziente Validierungsmöglichkeit. Der Experte würde einen minimalen Anteil in der vorgegeben Zeit zuordnen können. Daher wird dieser Ansatz nicht weiter verfolgt und im Rahmen eines teilautomatisierten Logistikzentrums ausgeschlossen.

Den zweiten Ansatz stellt die Validierung mit Hilfe einer Simulation dar, in Abschnitt 2.1.2.3 wurde bereits eine Validierung durch eine Simulation nach der Anwendung von DM vorgestellt und deren Anwendbarkeit bewiesen. Daher wird die Möglichkeit der Simulation bei dieser Problemstellung näher betrachtet. Eine Validierung durch die Simulation ist abhängig vom System leicht umsetzbar. Die zu untersuchenden Daten stammen in der Regel aus einem WMS und um Veränderung am WMS vorzunehmen, existiert ein Testsystem und in diesem Testsystem kann die Simulation durchgeführt werden. Um das Ergebnis aus den DM-Verfahren zu validieren ist es notwendig, diese Simulation durchzuführen. Eine andere Möglichkeit zu überprüfen, ob das System eine korrekte Zuordnung zu den beiden Lagerbereich ist derzeit nicht vorhanden. Im Bereich der Validierung muss ebenfalls eine der Restriktionen überprüft werden, denn das Ergebnis muss auf den maximal möglichen Durchsatz aus dem automatisierten Teil geprüft werden. Dabei muss die Simulation nach jeder Veränderung im Vorgehensmodell durchgeführt werden, um validierte Ergebnisse für den Lagerbereich geliefert zu bekommen. Die Validierung der Beispieldaten dieser Handlungsalternative wird aufgrund der Kapazität der Arbeit nicht durchgeführt. Um beide Handlungsalternativen miteinander vergleichen zu können, ist keine Validierung notwendig. Im abschließenden Vergleich wird lediglich der Aufwand der Durchführung der Validierung betrachtet.

4.3 Prognosewahrscheinlichkeit für die Auslagerung

Die zweite Handlungsalternative beinhaltet die Optimierung eines Parameters in einem Algorithmus. Deshalb ist es notwendig den vorhandenen Algorithmus zu kennen. Dieser Parameter soll eine Aussage darüber treffen, wie häufig das einzelne Produkt jeweils ausgelagert wird. Dies

steht im Widerspruch zu dem Titel dieser Arbeit, da die Einlagerung und nicht die Auslagerung optimiert werden soll. Um jedoch erfolgreich herauszufinden, wie viel in welchen Bereich eingelagert werden muss, muss vorher bestimmt werden, wie häufig ein Produkt ausgelagert wird. Da sich diese Häufigkeit auf die Bestellung des einzelnen Produktes durch den Kunden bezieht, wird diese Häufigkeit mit der Wahrscheinlichkeit der Auslagerung gleichgesetzt. In diesem Abschnitt wird beschrieben, wie die Wahrscheinlichkeit der Auslagerung mit geeigneten DM-Verfahren ermittelt werden kann. Dafür wird in Abbildung 26 der Ablauf dieses Abschnittes aufgezeigt.

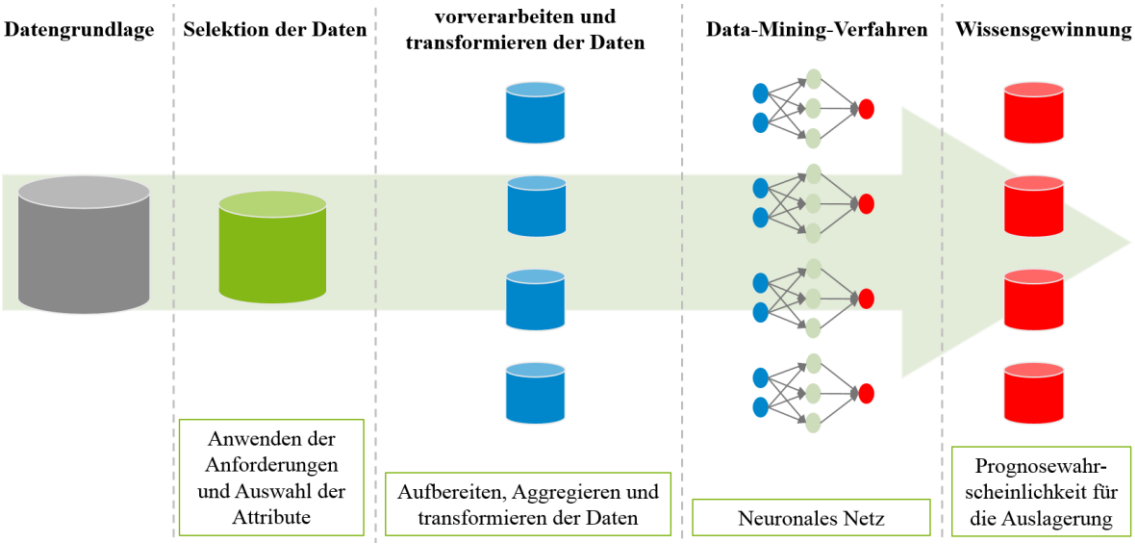


Abbildung 26: Vorgehensweise der Handlungsalternative Zwei

Bei der Entwicklung des Vorgehensmodells für die Handlungsalternative Zwei findet eine Orientierung an dem Vorgehensmodell von [FPS96] statt. Dabei sind die Phasen der Vorverarbeitung und Transformation der Daten in diesem Vorgehensmodell zusammengelegt. Die Grundlage bilden die bereits erläuterten Daten, wobei nur die Tabelle der Kundenbestellungen für dieses Vorgehensmodell genutzt wird. Nachdem die Restriktionen und Anforderungen auf die Daten angewendet wurden, können die Daten vorverarbeitet und transformiert werden. Dabei muss für jedes Produkt eine eigene Tabelle erstellt werden, weil zeitabhängige Daten vorliegen. Deshalb liegt für jedes Datum ein eigener Eintrag in der Tabelle vor. Mit Hilfe von diesen Tabellen werden die neuronalen Netze angelernt. Auf in der Zukunft liegenden Zeitreihen werden die gelernten Modelle angewendet. Damit wird eine zukünftige Wahrscheinlichkeit erzeugt, welche angibt, wie häufig die jeweiligen Produkte ausgelagert werden.

In dieser Handlungsalternative sind teilweise Überschneidungen im Bereich der Datenvorverarbeitung mit der Handlungsalternative Eins zu finden. Daher wird an überschneidenden Stellen auf die Handlungsalternative Eins verwiesen. Weiterhin findet keine Unterteilung des Abschnittes der Datenvorverarbeitung statt.

4.3.1 Vorverarbeitung der Daten

Eine Vorverarbeitung der Daten ist für das erfolgreiche Durchführen von DM-Verfahren notwendig. In Abbildung 27 wird der detaillierte Ablauf der Datenvorverarbeitung für diese Handlungsalternative beschrieben. Dabei wird mit der Datengrundlage begonnen. Die genutzten Daten sind bei dieser Handlungsalternative die Daten von Bestellungen der Kunden beim Logistikzent-

rum (Tabelle der Kundenbestellungen). Die Bestelldaten des Logistikzentrums werden nicht benötigt (Tabelle der Lieferantenbestellungen). In Abschnitt 4.2.1.1 werden beide Tabellen genutzt, um alle Produkte herauszufinden, welche sich derzeit im Lager befinden. Bei dieser Handlungsalternative werden nur die Produkte betrachtet, welche mindestens einmal ausgelagert wurden. Daher werden die Daten mit den Lieferantenbestellungen nicht betrachtet, da sie noch nie ausgelagert worden. Aus diesem Grund sind die Produkte der Lieferantentabelle nicht in der Kundetabelle vorhanden und werden nicht betrachtet. Bei einer Vorhersage für die Häufigkeit der Auslagerung der Produkte der Lieferantentabelle wäre der vorhergesagte Wert immer null.

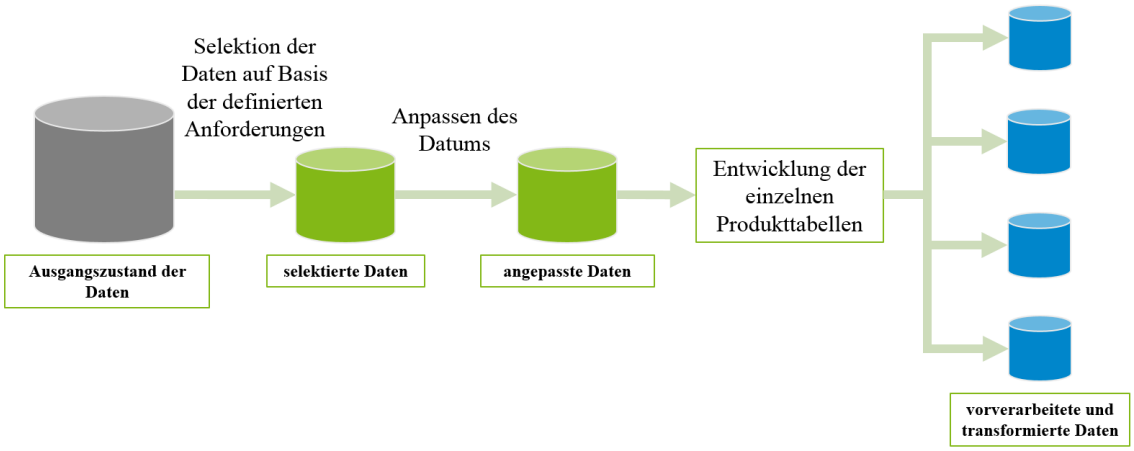


Abbildung 27: Vorverarbeitung der Daten in Handlungsalternative Zwei

Auf den Ausgangszustand der Daten wird im ersten Schritt eine *Selektion der Daten* durchgeführt, dabei werden definierte Anforderungen und Restriktionen angewendet. In diesem Schritt gibt es eine Überschneidung mit der ersten Handlungsalternative, deswegen wird auf Abschnitt 4.2.1.1 verwiesen. Für eine erfolgreiche Durchführung des DM-Verfahrens ist es notwendig, dass nur zeitabhängige Daten betrachtet werden. Somit müssen alle zeitunabhängigen Variablen entfernt werden. Dabei werden alle vertikalen Dimensionsreduktionen aus Handlungsalternative Eins ebenfalls in dieser Handlungsalternative vorgenommen. Bei der horizontalen Dimensionsreduktion werden nur zwei Restriktionen angewendet. Dabei werden die Daten selektiert, indem nicht betrachtete Typen von Bestellungen und nicht betrachtete Lagerbereiche herausgelöscht werden. Daraus entsteht eine Tabelle mit vier Attributen, welche eine Aussage über das Datum, die bestellte Menge, eine eindeutige ID für das Produkt und über die zugehörige Kategorie des Produktes trifft.

Nachdem die Selektion abgeschlossen ist, muss das *Datum angepasst* werden. Dabei muss das Datum als solches vom DM-Verfahren erkannt werden. Eine genaue Erläuterung von Möglichkeiten der Anpassung von Datumsangaben findet sich in Abschnitt 3.2.4. In diesem Schritt muss entschieden werden, in welchem Intervall die Daten angegeben werden. Dabei ist die ursprüngliche Ausgangsform zu berücksichtigen. Sollten die Daten tageweise vorliegen, sollten sie auch so verwendet werden. Damit wird die Notwendigkeit der Transformation von den gesamten Daten vermieden und es resultiert daraus eine erhebliche Zeitersparnis. Aus diesem Vorverarbeitungsschritt gehen als Ergebnis angepasste Daten hervor, welche im Folgenden weiter verwendet werden.

Mit den angepassten Daten können die Tabellen erstellt werden, welche das DM-Verfahren zur Vorhersage für die Auslagerung benötigt. Dieser Prozess ist komplex und wird daher in Abbildung 28 detailliert erläutert. Das Ziel dieses Prozesses besteht darin, dass für jedes Produkt eine eigene Tabelle erzeugt wird. In dieser Tabelle wird die Häufigkeit der Auslagerung des Produktes am jeweiligen Tag angegeben. Dabei ist die Auslagerung abhängig davon, welches Produkt wie häufig am Tag verschickt wurde. Die Grundlage der Daten sind die bereits angepassten Daten. Die angepassten Daten werden zur Berechnung von zwei verschiedenen Kennwerten genutzt.

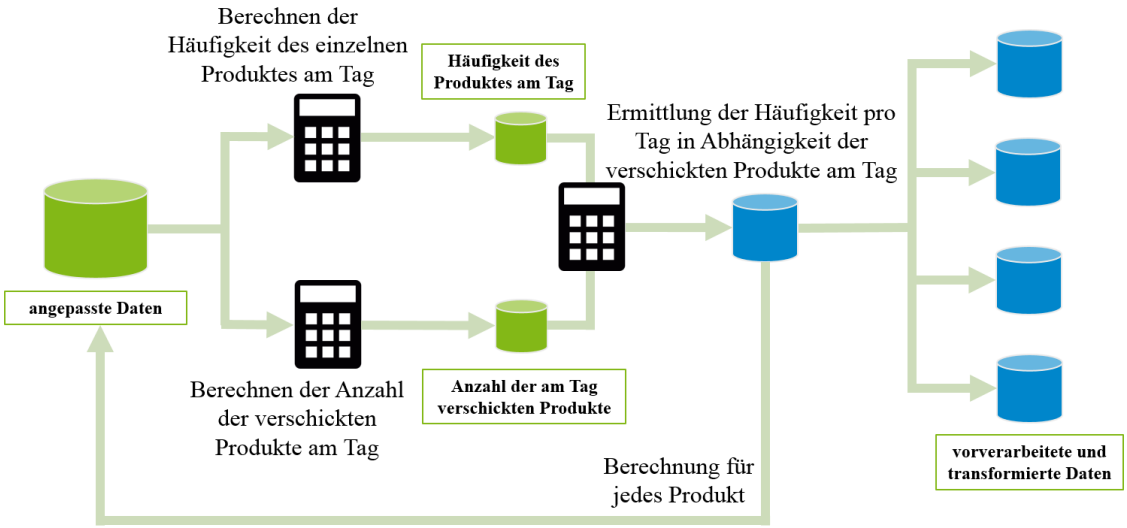


Abbildung 28: Ermittlung der zeitabhängigen Tabellen für Handlungsalternative Zwei

Der erste Kennwert beginnt mit der Berechnung *der Anzahl der verschickten Produkte am Tag*. Dafür werden alle Mengen an dem jeweiligen Tag summiert und dann in Abhängigkeit vom Datum in eine Tabelle geschrieben. Somit entsteht eine Tabelle mit zwei Attributen, das erste Attribut trifft eine Aussage über das fortlaufende Datum und das zweite Attribut über die Anzahl der verschickten Produkte am Tag. Die Abbildung 12 zeigt beispielhaft, wie ein solcher Verlauf der Zahlen aussehen kann. Das Ergebnis ist eine Tabelle mit der Anzahl der am Tag verschickten Produkte. Diese Tabelle verändert sich nicht, sondern bleibt bei jeder Berechnung für das einzelne Produkt gleich. Daher wird sie einmal berechnet und dann immer wieder der weiteren Berechnung zugeführt, um Rechnerkapazität zu sparen.

Bei dem zweiten Kennwert wird *die Anzahl der am Tag verschickten Produkte* berechnet. Auf Grundlage der angepassten Daten wird die Menge des einzelnen Produktes am jeweiligen Tag summiert und dann in Abhängigkeit vom Datum in eine Tabelle geschrieben. Dabei entsteht eine Tabelle mit zwei Attributen, das erste Attribut gibt die Auskunft über das Datum, bei dem zweiten Attribut sind die Mengen von einem Produkt am jeweiligen Tag festgeschrieben. Bei der hohen Anzahl an verschiedenen Produkten kann es passieren, dass ein Produkt an einem Tag nicht verschickt wird. Dadurch wird ein fehlender Wert erzeugt, welcher jedoch für die DM-Verfahren nicht zulässig ist, deswegen muss dieser fehlende Wert durch Null ersetzt werden. Diese Berechnung muss für jedes Produkt durchgeführt werden, dadurch entsteht für jedes Produkt eine eigene Tabelle mit den gleichen zeitabhängigen Attributen.

Mit diesen beiden Tabellen wird eine neue Tabelle generiert, welche die beiden neben dem Datum erzeugten Attribute verbinden soll. Dafür wird die *Häufigkeit des einzelnen Produktes am jeweiligen Tag* bestimmt. Hierfür wird für jedes Produkt ein neues Attribut generiert. Für die

Ermittlung des neuen Attributes wird der zweite Kennwert durch den ersten Kennwert geteilt. Mit dieser Berechnung wird die Häufigkeit der Auslagerung des Produktes bestimmt. Daraus entsteht eine Tabelle, bei der eine Häufigkeit des einzelnen Produktes in Abhängigkeit des jeweiligen Datums steht. Daraus entstehen Zeitreihen mit denen eine Anwendung von einem NN möglich ist.

Nach der Erstellung dieser Tabellen können weitere Attribute hinzugefügt werden. Dabei müssen die zugefügten Attribute zeitabhängig sein, um sie mit den erzeugten Tabellen verknüpfen zu können. Dabei kann ein solches Attribut eine Aussage darüber treffen, ob das Produkt besonders beworben wurde. Ebenfalls kann festgehalten werden, wie das Wetter in den betrachteten Zeitreihen war. Diese Attribute werden als externe Faktoren bezeichnet. Durch das Hinzufügen von diesen externen Faktoren wird das Lernen des NN im Bereich der DM-Verfahren wesentlich verbessert und dementsprechend auch die Leistungsfähigkeit des NN. Deswegen ist eine Berücksichtigung dieser externen Faktoren zu empfehlen.

Die Berechnung muss für jedes Produkt durchgeführt werden, somit entsteht eine hohe Anzahl an Tabellen. Daher ist eine weitere Dimensionsreduktion sinnvoll, indem nicht jedes einzelne Produkt betrachtet wird, sondern die Kategorie, der das Produkt zugehörig ist. Damit wird das Produkt durch die Kategorie ersetzt, wobei eine Reduktion der Dimension möglich ist. Die Algorithmen aus Abschnitt 2.1.2.2 erfordern jedoch produktabhängige Häufigkeitsvorhersagen. Deswegen muss eine weitere Berechnung zur Ermittlung der Häufigkeiten des einzelnen Produktes mit den Ergebnissen des DM durchgeführt werden, diese wird nach der Anwendung des DM-Verfahrens erläutert. Das hier entwickelte Modell bezieht sich immer auf die Produkte, weil eine Durchführung mit Produkten das beste Ergebnis liefert. Daher wird im Folgenden immer von Produkten gesprochen. Die Möglichkeit sie durch die Kategorie zu ersetzen, ist zu jederzeit möglich und stellt eine Art der Dimensionsreduktion dar.

4.3.2 Anwendung des Data-Mining-Verfahrens

Bei dieser Handlungsalternative wird das NN als DM-Verfahren betrachtet. Das NN eignet sich gut, um eine Vorhersage von Zeitreihen zu treffen. Als Grundlage werden die vorverarbeiteten und transformierten Daten aus dem vorherigen Abschnitt verwendet. In der Abbildung 29 sind symbolhaft vier Tabellen dargestellt, wie bereits erläutert liegt für jedes Produkt eine eigene Tabelle mit der Häufigkeit der Auslagerung vor.

Jede dieser Tabellen wird mit einem DM-Verfahren bearbeitet, in diesem Fall dem NN. Somit wird im Bereich der *Entwicklung des NN* der Prozess zum Erlernen eines DM-Modells durchgeführt. Der Prozess des Erlernens des DM-Modells ist dem aus Abbildung 22 gleich. Der Unterschied der beiden Handlungsalternativen besteht darin, dass in dieser Handlungsalternative nur ein DM-Verfahren angewendet wird und kein Vergleich von verschiedenen Verfahren stattfindet. Der Ablauf unterscheidet sich in keiner Hinsicht. Daher wird für jedes Produkt der Prozess in Abbildung 22 durchgeführt und es entsteht für jedes Produkt ein eigenes NN. Dieses validierte Modell wird genutzt, um es auf den neuen Daten anzuwenden.

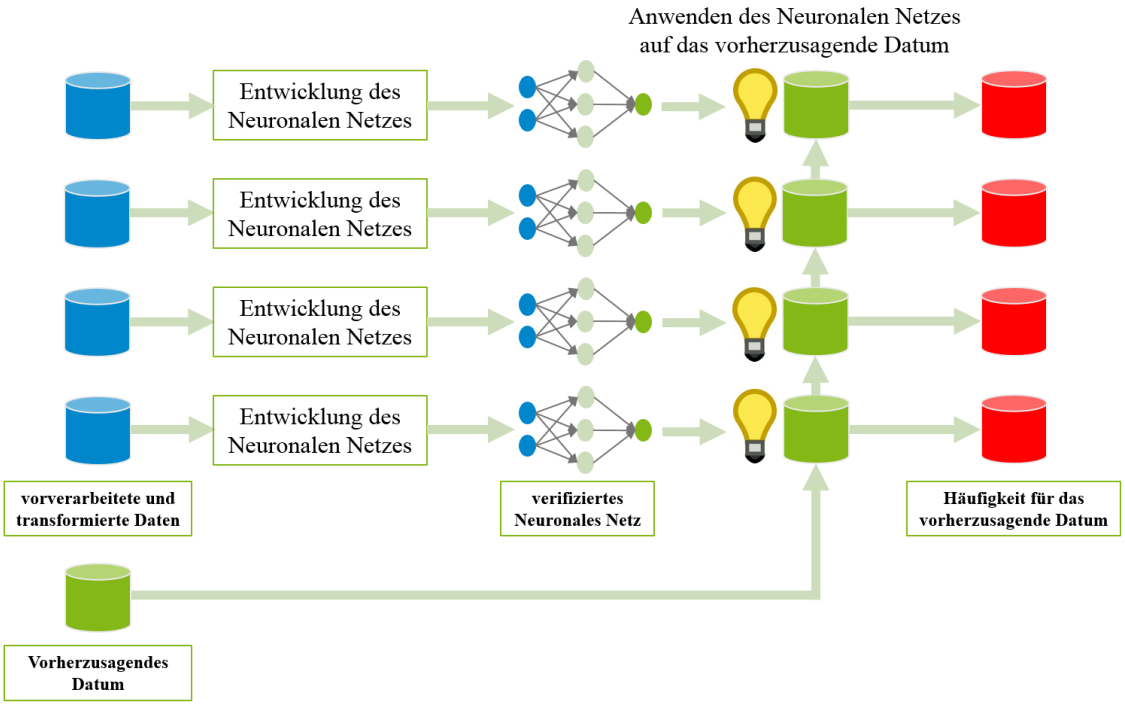


Abbildung 29: Ablauf zur Anwendung des Data-Mining-Verfahrens in Handlungsalternative Zwei

Der nächste Prozessschritt beschäftigt sich mit dem *Anwenden des NN auf dem vorzusagenden Datum*. Dazu wird eine neue Tabelle erstellt, in der die vorherzusagenden Daten abgebildet sind. Über die neue Tabelle wird ebenfalls der Prognosezeitraum bestimmt, eine Wahl von mehr als vier Wochen ist nicht zu empfehlen. Der Prognosezeitraum ist abhängig von der Menge der vorhandenen Vergangenheitsdaten. Je mehr Vergangenheitsdaten vorliegen, desto länger kann der Prognosezeitraum sein und umso genauer ist die Prognose. Dabei existieren jedoch Grenzen, so sollten Vergangenheitsdaten nicht betrachtet werden, die älter als drei Jahre sind. Dies hängt mit den Bestellungen der Kunden beim Logistikzentrum zusammen, denn in einem teilautomatisierten Logistikzentrum liegt ein hoher Wechsel der bestellten Produkte vor. Daher muss der Produktlebenszyklus eines Produktes im Logistikzentrum berücksichtigt werden, um einen erfolgreichen Prognosezeitraum zu ermitteln. Eine Berücksichtigung von externen Faktoren muss in diesem Prozessschritt ebenfalls stattfinden. Die externen Faktoren sollten bei den zu prognostizierenden Daten bereits eingetragen sein, sofern welche vorhanden sind. Das Modell des NN wird auf diese neuen Daten angewendet und liefert somit die zukünftige Wahrscheinlichkeit der Auslagerung für die jeweiligen Produkte. Mit diesem Ergebnis kann nun der Algorithmus zur Lagerplatzvergabe unterstützt werden.

4.3.3 Transformation der Ergebnisse

Eine Transformation der Ergebnisse ist notwendig, wenn die Kategorie genutzt wurde, um eine Ermittlung die Wahrscheinlichkeit für die Auslagerung für jedes einzelne Produkt zu bekommen. Der Prozess zur Rücktransformation gliedert sich in zwei Teilprozesse auf, in Abbildung 30 ist der erste Teilprozess zu erkennen. Dabei wird im ersten Teilschritt berechnet, mit welcher Häufigkeit ein Produkt in einer Kategorie vorkommt. Als Grundlage dienen die angepassten Daten, welche identisch mit denen aus der Datenvorverarbeitung sind. Die angepassten Daten werden auf zwei verschiedene Berechnungen aufgeteilt. Dabei werden die Produkte gezählt und die

Anzahl der Produkte in einer Kategorie bestimmt. Beim Zählen der Produkte wird die Anzahl der Produkte in der angepassten Tabelle betrachtet, somit werden die IDs der Produkte gezählt und in Abhängigkeit von dieser ID in eine Tabelle geschrieben. Bei der Berechnung der Anzahl der Produkte in einer Kategorie werden ebenfalls die eindeutigen IDs der Produkte gezählt, doch bei dieser Berechnung in Abhängigkeit von der Kategorie. Aus diesen beiden Berechnungen entstehen zwei Tabellen, eine mit der Anzahl der Produkte je Produkt und eine mit der Anzahl der Produkte in einer Kategorie.

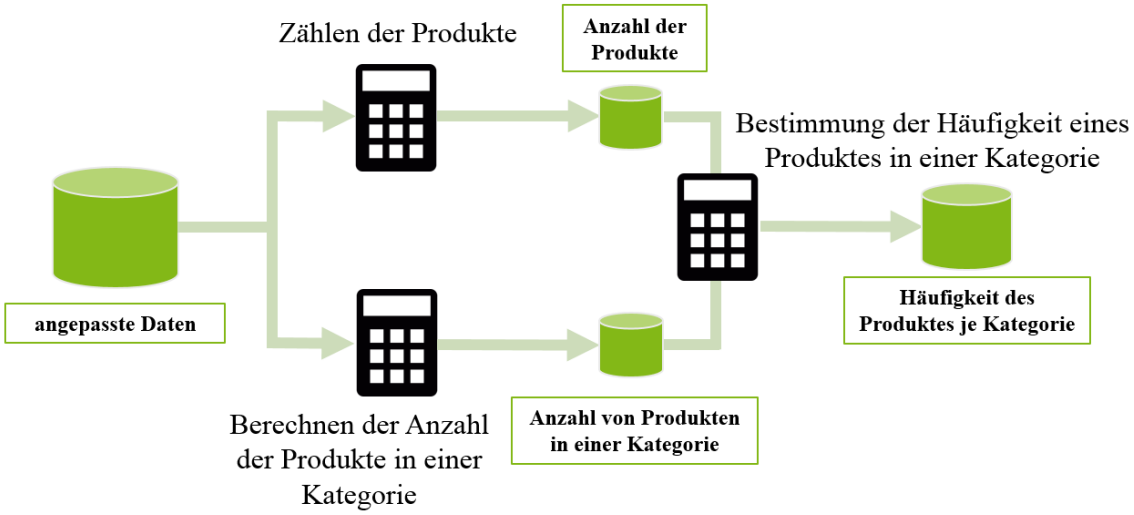


Abbildung 30: Erster Teilschritt zur Rücktransformation der Daten in Handlungsalternative Zwei

Dadurch kann jetzt die Häufigkeit der einzelnen Produkte in der jeweiligen Kategorie ermittelt werden. Indem die Anzahl der Produkte durch die Anzahl von Produkten in einer Kategorie geteilt wird. Aus dieser Berechnung resultiert eine Tabelle, welche eine Aussage über die Häufigkeit des Produktes je Kategorie trifft. Diese Tabelle beinhaltet drei Attribute, das erste Attribut beschreibt die Kategorie, das zweite Attribut die eindeutige ID des Produktes und das dritte Attribut die Häufigkeit des Produktes in der jeweiligen Kategorie. Die Tabelle wird im zweiten Teilprozess benötigt, welcher in Abbildung 31 dargestellt ist.

Dieser Teilschritt benötigt zwei Schleifen, um eine erfolgreiche Ermittlung der Häufigkeiten für die Produkte durchführen zu können. Als Grundlage wird die ermittelte Häufigkeit des Produktes je Kategorie genutzt. In dieser Tabelle wird als erstes das Attribut der Kategorie benötigt, um eine Aufteilung der Tabellen je Kategorie durchführen zu können. Das Ergebnis des DM liegt je Kategorie vor, daher müssen jeweils die Kategorien mit ihren zugehörigen Produkten einzeln betrachtet werden. Die Tabelle der Häufigkeit des Produktes je Kategorie wird somit in alle Kategorien aufgeteilt. Für jede Kategorie wird eine eigene Tabelle mit drei Attributen erzeugt. In der Abbildung 31 hat eine dieser erzeugten Tabellen den Namen Produkte der Kategorie C. In diesen Tabellen sind alle zugehörigen Produkte einer Kategorie mit ihrer jeweiligen Häufigkeit vorhanden.

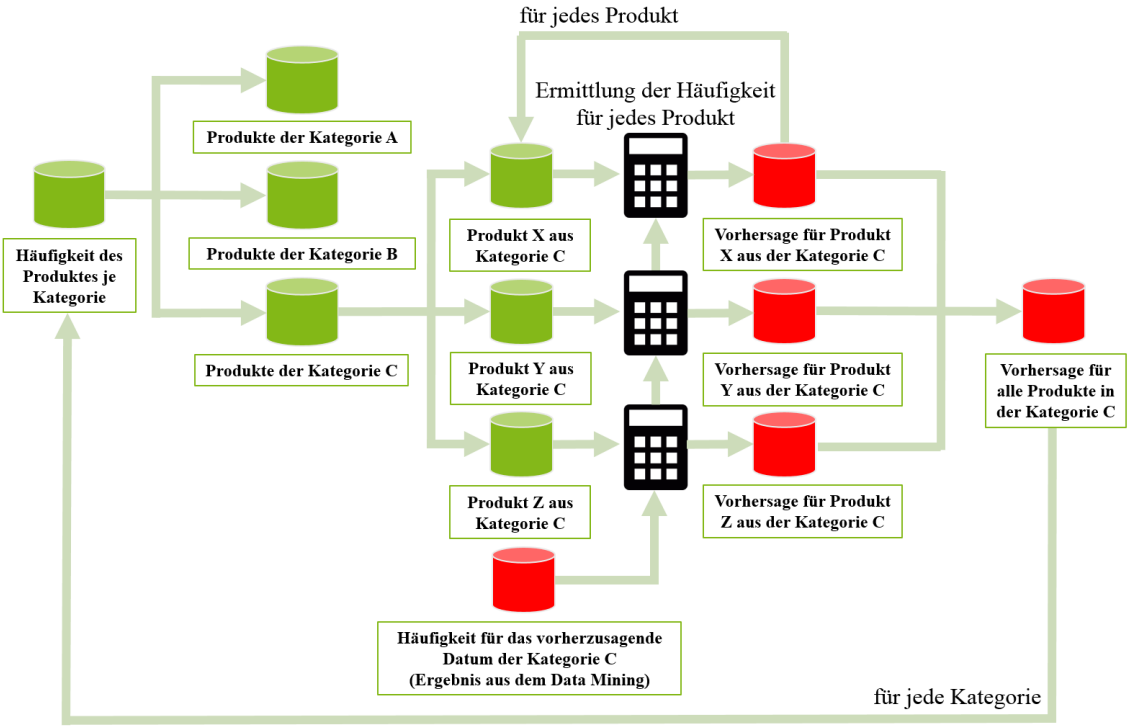


Abbildung 31: Zweiter Teilschritt zur Rücktransformation der Daten in Handlungsalternative Zwei

Die Tabelle mit dem Namen Produkte der Kategorie C wird weiter betrachtet. Dafür wird sie im ersten Schritt der Berechnung zur *Ermittlung der Häufigkeit für jedes Produkt* zugeführt. Neben dieser Tabelle wird ebenfalls das Ergebnis der Kategorie C aus dem DM der Berechnung hinzugefügt. Dementsprechend wird aus der Häufigkeit für das vorherzusagende Datum der Kategorie C und den Produkten der Kategorie C berechnet, wie häufig das einzelne Produkt in der Zukunft ausgelagert wird. Dafür wird der vorhergesagte Wert aus dem DM mit der Häufigkeit der Produkte in der Kategorie C multipliziert. Als Ergebnis wird für jedes Produkt eine eigene Tabelle erzeugt. Diese beinhaltet zwei Attribute, die Angabe des Datums und die Wahrscheinlichkeit zur Auslagerung des einzelnen Produktes. Der beschriebene Prozessschritt muss für alle Produkte durchgeführt werden, welche in der Kategorie C sind.

Wenn die Berechnung für alle Produkte durchgeführt wurde, können alle Tabellen zu einer Tabelle zusammengeführt werden. In der entstandenen Tabelle sind alle Produkte einer Kategorie abgebildet. Somit existieren unterschiedlich viele Attribute, immer abhängig davon, wie viele Produkte es in einer Kategorie gibt. Für jedes Produkt wird ein eigenes Attribut mit der Aussage über die Wahrscheinlichkeit der Auslagerung angelegt. Das Attribut der Datumsangabe verändert sich nicht, weil von diesem Attribut Angaben zur Wahrscheinlichkeit der Auslagerung abhängig sind. Nachdem dies für eine der Kategorien abgeschlossen ist, muss dies für alle vorhandenen Kategorien durchgeführt werden. Daher werden als nächstes die Produkte der Kategorie B betrachtet und es findet die Berechnung für jedes Produkt der Kategorie B statt. Als Ergebnis des Prozesses existieren so viele Tabellen, wie auch Attribute vorliegen. Die Verbesserung besteht darin, dass in jeder Tabelle die Wahrscheinlichkeit der Produkte einzeln dargestellt wird und somit eine genauere Vorhersage möglich ist. Mit der Transformation der Ergebnisse ist das Vorgehensmodell beendet.

4.3.4 Validierung des Vorgehensmodells

Für die Validierung der zweiten Handlungsalternative steht nur eine Möglichkeit zur Verfügung. Mit der Anwendung des NN wurde eine Zeitreihenprognose durchgeführt und diese wird mit statischen Prognoseberechnung validiert. Dabei muss berücksichtigt werden, dass die Verifizierung des NN bereits bei der Anwendung des DM-Verfahrens stattgefunden hat (vgl. Abschnitt 4.2.2). In diesem Abschnitt wird eine Validierung des Modells mit Hilfe von Prognoseverfahren durchgeführt. Dafür werden zukünftige Häufigkeiten der Auslagerung prognostiziert. In der Literatur existieren mehrere Verfahren zur Prognoseberechnung. In dieser Arbeit wird das Arithmetische Mittel aufgrund seiner einfachen Anwendbarkeit verwendet. Für vertiefende Literatur wird in diesem Fall auf [Sch94] verwiesen, aus dieser Literatur stammt die angewendete Formel. Zum Vergleich werden die beiden am meisten verkauften Produkte der Beispieldaten herausgesucht und das arithmetische Mittel ausgerechnet. Das arithmetische Mittel berechnet sich nach folgender Formel [Sch94]:

$$M_t = \frac{x_t + x_{t-1} + x_{t-2} + \cdots + x_{t-N+1}}{N}$$

Dabei gilt:

M_t = arithmetische Mittel

X_t = Menge zum Zeitpunkt t

N = Anzahl an untersuchten Mengen

T = Periode

Zur Berechnung der Prognose werden Vergangenheitsdaten aus dem letzten Jahr genommen, daher sind die Diagramme sehr groß. Zur besseren Veranschaulichung befinden sich diese Diagramme im Anhang A4 (Produkt 33912122) und A5 (Produkt 36435968). Die zugrunde liegenden Daten befinden sich im elektronischen Anhang EA10. In diesem Abschnitt wird ein Ausschnitt aus den letzten 30 Tagen und den zukünftigen 28 Tagen gezeigt. Dabei muss berücksichtigtet werden, dass die Grundlage für die Berechnung der Prognose den vollen Zeitraum des letzten Jahres umfasst.

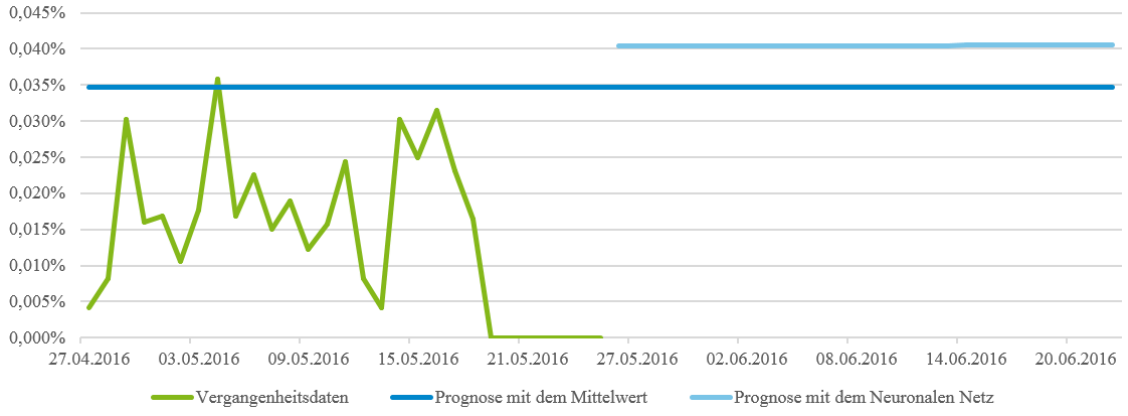


Abbildung 32: Prognosevergleich für das Produkt 33912122

In der Abbildung 32 ist der Prognoseverlauf für das Produkt 33912122 zu erkennen. Aufgrund von internen sensiblen Unternehmensdaten wird mit eindeutigen Nummern und nicht mit den Namen gearbeitet. Insgesamt hat die Abbildung 32 drei verschiedene Graphen, der grüne Graph stellt die Vergangenheitsdaten dar und die beiden blauen stellen jeweils die Prognosewerte dar.

An den Vergangenheitsdaten ist zu erkennen, dass eine hohe Schwankung im Abverkauf des Produktes vorlag. Ab dem 21.05.2016 wurde das Produkt überhaupt nicht mehr verkauft. Die dunkelblaue Kurve zeigt die Prognose mit dem Mittelwert und die hellblaue die Prognose mit dem NN. Dabei ist zu erkennen, dass die Prognose des NN über der des Mittelwertes liegt. Um das erklären müssen sich die gesamten Daten der Vergangenheit angesehen werden. In diesen Daten (Anhang A4) ist zu erkennen, dass erhebliche Schwankungen in den Vergangenheitsdaten vorliegen. Die Ausprägung der maximalen Werte ist wesentlich höher als in Abbildung 32. Die Prognosewerte für das NN und den Mittelwert unterscheiden sich nur um 0,005%, dies ist marginal und daher ist eine Validierung des NN mit Hilfe des Mittelwertes möglich.

Ein zweites Produkt wird in der Abbildung 33 betrachtet, die Zuordnung der einzelnen Graphen verhält sich dem vorherigen Abschnitt gleich. Bei dem grünen Graph werden die Vergangenheitsdaten dargestellt. Dabei ist zu erkennen, dass am Anfang des Zeitraumes keine Produkte ausgelagert wurden und dieser somit 0 beträgt. Zwischen dem 03.05.2016 und 09.05.2016 finden wieder Auslagerungen statt. Die folgenden Werte schwanken nicht so stark, im Vergleich zur vorherigen Abbildung. Der dunkelblaue Graph zeigt die Prognose mit dem Mittelwert, dieser Graph liegt wesentlich höher als der hellblaue Graph. Dabei handelt es sich um einen Unterschied von knapp 0,2%, was einen hohen Unterschied darstellt in Berücksichtigung der lokalen Maxima und Minima. Um ein besseres Verständnis zu erlangen, werden die gesamten Vergangenheitsdaten in Anhang A5 betrachtet. Dabei ist zu erkennen, dass lediglich in der Hochphase des Logistikzentrums viel von diesem Produkt verkauft wurde. Im Vergleich dazu, über den Rest des Jahres wesentlich weniger. Da die Prognose über den Mittelwert genau solche Schwankungen nicht berücksichtigt [Sch94], wird angenommen, dass die Prognose des Mittelwertes zu hoch angesetzt ist. Ebenfalls muss bei der Betrachtung der Prognose des NN berücksichtigt werden, dass durch die Transformation von der Kategorie zum einzelnen Produkt Fehler entstanden sein können und daraus Abweichungen resultieren. Eine Validierung der Prognose des NN dieses Produktes mit Hilfe des Mittelwertes ist nicht möglich. Durch die beschriebenen Einflüsse besteht jedoch die Möglichkeit, dass die Prognose des NN korrekt ist.

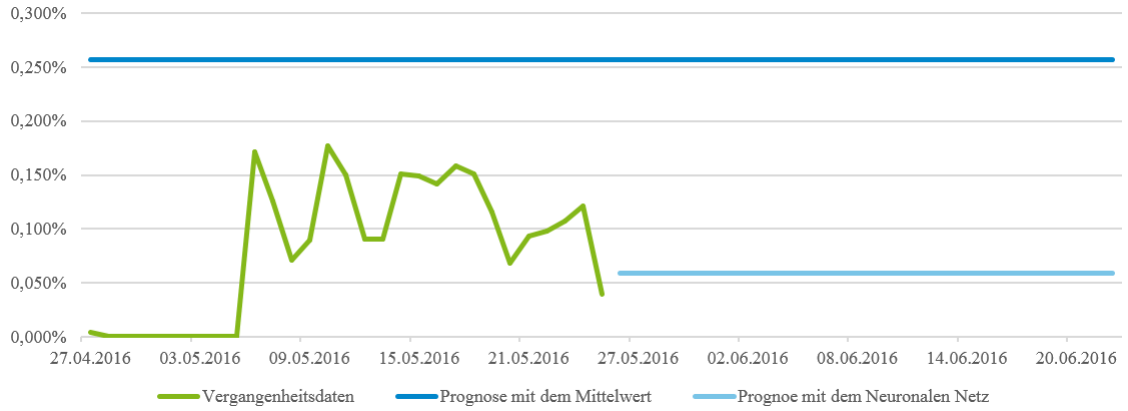


Abbildung 33: Prognosevergleich für das Produkt 36435968

Abschließend muss erwähnt werden, dass lediglich zwei verschiedene Produkte von über 100.000 angeschaut wurden. Die testweise Validierung war erfolgreich, jedoch muss in Anbetracht der großen Datenmenge überlegt werden, ob eine einmalige Simulation der Ergebnisse durchgeführt wird. Mit der Simulation können alle Ergebnisse und nicht nur ein Teil validiert

werden. Dabei sollte in der Simulation die Wahrscheinlichkeit der Auslagerung in einen Algorithmus aus Abschnitt 2.1.2.2 oder ein selber entwickelten Algorithmus eingefügt werden. Die Simulation muss nur einmal durchgeführt werden, nicht bei jedem neuen Trainieren der NN.

4.4 Vergleich der beiden Handlungsalternativen

In diesem Abschnitt werden die beiden Handlungsalternativen auf der Basis von verschiedenen Einflussgrößen miteinander verglichen. Dabei werden verschiedene Einflussgrößen definiert, welche sich aus einer praktischen und theoretischen Sichtweise aus den beiden Handlungsalternativen ableiten. Am Ende des Abschnittes wird eine Entscheidung für eine der beiden Handlungsalternativen getroffen. Dabei werden keine quantitativen, sondern nur qualitative Einflussgrößen verwendet. Die Einflussgrößen lauten wie folgt:

- Validierung
- Implementierung
- Datengrundlage
- Transparenz der Modellergebnisse
- Benötigte Rechenleistung
- Häufigkeit der Durchführung

Die Einflussgröße zur Betrachtung der *Validierung* stellt den ersten Vergleich dar. Zu beiden Handlungsalternativen wurden Möglichkeiten zur Validierung aufgezeigt. Dabei ist in Handlungsalternative Eins eine Validierung nur über die Simulation möglich. Dadurch werden viele Simulationsdurchläufe benötigt, um ein validiertes Ergebnis zu erhalten. Durch die mehrmalige Simulation der Ergebnisse wird daraus ein iterativer Prozess zwischen der Simulation und dem Vorgehensmodell. Das ist im Vergleich zur zweiten Handlungsalternative zeitaufwändig und rechenintensiv. Bei der zweiten Handlungsalternative stehen zwei Verfahren zur Validierung zur Verfügung, eine Validierung über Prognoseverfahren oder eine Validierung über die Simulation. Die Verfahren zur Bestimmung der Prognose auf Zeitreihen sind weit verbreitet und lassen sich daher einfach anwenden und benötigten im Vergleich zur Simulation kein Expertenwissen. Mit den Prognoseverfahren lässt sich der ermittelte Wert der NN validieren. Um jedoch eine Validierung des gesamten Algorithmus durchführen zu können, ist eine Simulation notwendig. Diese Simulation erfordert nicht die Anzahl der Durchläufe, die bei Handlungsalternative Eins benötigt werden und sie muss nicht iterativ gestaltet werden. Dadurch lässt sich bei dieser Einflussgröße festhalten, dass die Validierung bei Handlungsalternative Zwei nicht so zeitintensiv und komplex ist, wie in Handlungsalternative Eins. Daher ist in dieser Einflussgröße Handlungsalternative Zwei vorzuziehen.

Die zweite Einflussgröße untersucht die *Implementierung* in die bestehende Software. Insbesondere soll dabei der Programmieraufwand betrachtet werden, welcher eine Änderung der bestehen Einlagerungsstrategie verursacht. Bei der ersten Handlungsalternative muss neben dem Vorgehensmodell eine Möglichkeit zu Validierung programmiert werden, da die Simulation ein fester Bestandteil des Vorgehensmodells ist. Durch den Vergleich der Schritte des Vorgehensmodells der beiden Handlungsalternativen würde der Aufwand der Implementierung bei beiden Handlungsalternativen ungefähr gleich hoch sein. Die Vorverarbeitung der Handlungsalternative Zwei ist geringer als bei der ersten, jedoch muss bei der zweiten Handlungsalternative am Ende

eine Transformation der Daten durchgeführt werden. Dadurch gestaltet sich der Zeitaufwand bei beiden Handlungsalternativen ungefähr gleich. Bei der zweiten Handlungsalternative wird nur ein Parameter für einen Algorithmus entwickelt, daher muss der Algorithmus ebenfalls implementiert werden. Hierdurch entsteht bei Handlungsalternative Zwei ein höherer Programmieraufwand in diesem Bereich. Letztendlich müssen beide Handlungsalternativen soweit implementiert werden, dass sie als Ergebnis jeweils einen Lagerbereich angeben. Bei dieser ganzheitlichen Betrachtung wird die notwendige Implementierung einer Simulation aus Handlungsalternative Eins kompensiert. Daher kann für diese Einflussgröße keine Handlungsalternative eindeutig vorgezogen werden.

Die benötigten Daten zur Durchführung des DM-Modells unterscheiden sich in beiden Handlungsalternativen, deshalb wird die Einflussgröße der *Datengrundlage* näher erläutert. Bei der ersten Handlungsalternative werden globale Trainingsdaten für das jeweilige DM-Verfahren benötigt, welche den korrekten Lagerbereich voraussagen. Diese globalen Trainingsdaten müssen repräsentativ gegenüber der gesamten Anzahl an einzulagernden Produkten sein. Dabei kann es sich als Problem herausstellen, diese Trainingsdaten zu erlangen. Das stellt eine Schwachstelle in der ersten Handlungsalternative dar, welche in der zweiten Handlungsalternative nicht vorhanden ist. Für die zweite Handlungsalternative werden nur die Daten aus der Vergangenheit benötigt, somit sind keine globalen Trainingsdaten notwendig. Bei dieser Einflussgröße wird sich für die Handlungsalternative Zwei entschieden, da sie mit den bestehenden Daten des Systems eine Lösung findet.

Bei der *Transparenz* wird die Nachvollziehbarkeit der Ergebnisse untersucht. Da diese Einflussgröße in der Entwicklung von IT-Systemen immer bedeutender wird, wird ihm ein höherer Stellenwert beigemessen. Die erste Handlungsalternative hat als Ergebnis die direkte Zuordnung zu dem jeweiligen Lagerbereich, diese direkte Entscheidung ist für den Anwender nur schwer nachvollzierbar. Die Entscheidung beruht auf statistischen Verfahren, welche nur das Ergebnis präsentieren. Im Gegensatz dazu ermöglicht die Handlungsalternative Zwei einen besseren Einblick in die Entscheidung für den jeweiligen Bereich. Das DM wird eingesetzt um einen Wert zur Wahrscheinlichkeit der Auslagerung zu bestimmen, welcher durch die Validierung mit Prognoseverfahren nachvollziehbar ist. Ein eigener Algorithmus berücksichtigt die Restriktionen, welche im DM nicht abgedeckt werden können. Dazu zählt vor allem die Überprüfung des maximal möglichen Durchsatzes aus dem automatisierten Teil, welche im Algorithmus mit eingebracht werden kann. Die höhere Transparenz und Nachvollziehbarkeit liegt bei Handlungsalternative Zwei vor, weil Handlungsalternative Eins ein nur schwer nachvollziehbares Ergebnis liefert. Aus den genannten Gründen wird in diesem Fall eine Entscheidung mit einer höheren Gewichtung zugunsten der Handlungsalternative Zwei getroffen.

Bei dem Vergleich der benötigten *Rechnerleistung* wird auf die Durchführung mit den Beispieldaten zurückgegriffen, daher sind diese ein grober Richtwert. Eine Verbesserung kann durch ein leistungsstärkeres Rechnersystem erreicht werden, dabei verringert sich die Durchlaufzeit bei allen Verfahren. Diese Einflussgröße kann nur betrachtet werden, wenn beide Handlungsalternativen mit den gleichen Ressourcen durchgeführt wurde. Bei der ersten Handlungsalternative ist die Berechnung des Lagerbereiches mit dem Entscheidungsbaum und dem NN im Vergleich zur SVM schnell durchgeführt. Dabei benötigen das NN und der Entscheidungsbaum jeweils ungefähr 30 Minuten zur Ermittlung des Ergebnisses und die SVM ungefähr sieben Tage. Bei der

zweiten Handlungsalternative ist das davon abhängig, ob das für jedes Produkt berechnet wird oder für eine übergeordnete Kategorie. Bei der Berechnung für jedes Produkt hätte das mit den Beispieldaten ungefähr 21 Tage gedauert, bei der Nutzung der übergeordneten Kategorie ist dieser Wert mit der Rücktransformation auf ungefähr drei Stunden gesunken. Neben der Zeit für die Durchführung des Vorgehensmodells muss die Zeit für die Validierung mit einbezogen werden. Dabei benötigt die zweite Handlungsalternative weniger Zeit, als die erste Handlungsalternative. Dadurch fällt die kurze Dauer des Entscheidungsbaumes und des NN weniger in das Gewicht, denn die zeitintensive Validierung lässt diesen Vorteil verschwinden. Bei dieser Einflussgröße wird sich für die zweite Handlungsalternative entschieden, Dabei muss die Rechenzeit des Algorithmus berücksichtigt werden, in den die Wahrscheinlichkeit der Auslagerung implementiert wird. Dies muss an dieser Stelle aufgrund der eingegrenzten Problemstellung der Arbeit vernachlässigt werden.

Die letzte zu vergleichende Einflussgröße stellt die *Häufigkeit der Durchführung* der beiden Vorgehensmodelle zur Ermittlung des Lagerbereichs dar. Diese Einflussgröße ist stark abhängig vom teilautomatisierten Logistikzentrum und seinem Durchsatz. Eine Betrachtung muss trotzdem stattfinden, um eine Entscheidung für eine Handlungsalternative treffen zu können. In diesem Fall wird sich auf den beispielhaften Durchsatz in Abbildung 12 bezogen. Für beide Handlungsalternativen muss einmal in der Woche das Vorgehensmodell durchgeführt werden. Dabei muss bei der Handlungsalternative Eins, mit jedem Durchführen des Vorgehensmodells die Validierung durch eine Simulation durchgeführt werden. Dies ist zeitaufwendig, wie in den vorhergehenden Einflussgrößen bereits beschrieben. Bei der Handlungsalternative Zwei muss die Simulation nicht bei jeder Durchführung des Vorgehensmodells herangezogen werden, sondern ist nur bei der ersten Einführung des Vorgehensmodells notwendig. Für die weiteren Validierungen reicht ein Vergleich mit den erwähnten Prognoseverfahren. Die Häufigkeit der Durchführungen ist bei beiden Handlungsalternativen gleich, daher ist keine Entscheidung für eine der beiden Handlungsalternativen möglich. Eine Berücksichtigung des Zeitaufwandes für die erste Handlungsalternative sollte erfolgen.

Nach der Erläuterung der sechs Einflussfaktoren wird eine Entscheidung für die zweite Handlungsalternative getroffen, dies hängt insbesondere mit der besseren Transparenz des Verfahrens zusammen. Neben der Transparenz steht der benötigte Zeitaufwand bei der wöchentlichen Durchführung im Vordergrund, mit der zweiten Handlungsalternative ist wesentlich weniger Rechenleistung notwendig. Die Datengrundlage ist für die Durchführung der Handlungsalternative Zwei leichter zu beschaffen. Die globalen Trainingsdaten für die erste Handlungsalternative sind schwer zu bekommen, diese liegen entweder vor oder müssen mit Verfahren des unüberwachten Lernens ermittelt werden. Nachteilig bei der zweiten Handlungsalternative ist die Implementierung und die Suche und Entwicklung nach einem Algorithmus, welcher den Lagerbereich bestimmt. Dafür existieren in der Literatur mögliche Lösungsvorschläge, weswegen das Problem in dieser Arbeit vernachlässigt werden kann. Für beide Möglichkeiten muss eine Funktionsweise vor der Implementierung genauestens geprüft werden. Zur Veranschaulichung der zweiten Handlungsalternative, wird das Vorgehensmodell prototypisch mit Rapidminer umgesetzt.

5 Prototypische Umsetzung des entwickelten KDD-Vorgehensmodells

In diesem Kapitel wird das entwickelte Vorgehensmodell der zweiten Handlungsalternative prototypisch umgesetzt. Dafür werden die Daten eines konkreten Beispiels genutzt, welche teilweise im vorherigen Kapitel erwähnt wurden. In diesem Beispiel handelt es sich um ein teilautomatisiertes Logistikzentrum, welches im Onlinehandel tätig ist. Dieses teilautomatisierte Logistikzentrum wurde von der SSI Schäfer Noell GmbH als Generalunternehmer errichtet und aus diesem stammen die verwendeten Daten. Eine kurze Vorstellung der SSI Schäfer Noell GmbH erfolgt zu Beginn des Kapitels

Um die prototypische Umsetzung durchzuführen, wird das Programm Rapidminer verwendet. Auf Basis der Vorstellung von Rapidminer wird die zweite Handlungsalternative nach dem im vorherigen Kapitel beschriebenen Vorgehen umgesetzt. Beim Lesen dieses Kapitels wird ein Grundverständnis des Lesers für das Programm Rapidminer empfohlen. Ebenfalls ist ein Grundverständnis für Logiken in der Programmierung notwendig, um alle Prozesse nachvollziehen zu können. Sollte der Leser noch nicht ausreichende Vorkenntnisse über Rapidminer haben, wird empfohlen die von Rapidminer bereitgestellten Tutorials vorher durchzuführen. Zur Handhabung der Daten wurden zwei Programme benötigt. Da die der Kundenbestellungen als .csv Datei vorliegen und über 7.000.000 Einträge haben, können sie nicht direkt in Rapidminer eingelesen werden. Deswegen wurde ein CSV-Splitter genutzt, um die Daten in sieben gleich große Teile aufzuteilen. Die Daten wurden einzeln eingelesen in Rapidminer und nach der Selektion von Attributen zu einer Tabelle mit 71 Attributen zusammengefasst. Dabei wurden alle Attribute entfernt, welche im Anhang A3 als datenbankspezifisch bezeichnet wurden.

5.1 Vorstellung SSI Schäfer Noell GmbH

Zur besseren Einordnung der Arbeit wird im folgenden Teil kurz die Firmengruppe SSI Schäfer, insbesondere jedoch der Intralogistikhersteller SSI Schäfer Noell GmbH vorgestellt. SSI (Schäfer Systems International) Schäfer ist der weltweit führende Anbieter von Lager und Logistiksystemen und gliedert sich in die Firmen SSI Schäfer/Fritz Schäfer GmbH, SSI Schäfer Noell GmbH, SSI Schäfer Peem GmbH und Salomon Automation GmbH.

Die SSI Schäfer Noell GmbH realisiert komplexe Logistiksysteme. Der Aufgabenbereich erstreckt sich von der Systemplanung und -beratung bis hin zur schlüsselfertigen Anlage. Neben der Planung werden ebenfalls die IT-Lösungen mit eigenen Standards bzw. auf Basis von SAP-Technologie geliefert. Das Portfolio erstreckt sich über Hochregallager für Paletten, bis hin zu kompletten Kommissioniersystemen für den pharmazeutischen Bereich.

5.2 Umsetzung in Rapidminer

In diesem Abschnitt wird die Handlungsalternative in Rapidminer umgesetzt, da viele Schritte notwendig sind besteht dieser Abschnitt aus Unterabschnitten. Ebenfalls sind in diesen einzelnen

Unterabschnitten teilweise mehrere Prozesse aus Rapidminer dargestellt. Dafür wird im ersten Unterabschnitt die Vorverarbeitung beschrieben, im darauffolgenden das DM-Verfahren und beendet wird dieser Abschnitt mit der Transformation der Daten. Die jeweilig abgebildeten Operatoren besitzen immer unterschiedliche Parameter, welche vom Anwender festzulegen sind. Im Folgenden werden die einzustellenden Parameter teilweise genauer erläutert, sofern sie von dem ursprünglichen Wert abweichen. Neben der bildlichen Darstellung in dieser Arbeit, befinden sich die Prozesse im elektronischen Anhang EA1 – EA9. Dafür befindet sich in Verzeichnis im Anhang A6. Dieser besitzt die gleiche Notation, wie die Abbildungen in den nachfolgenden Abschnitten. Die vorgestellten und hinterlegten Prozesse wurden mit der Version 7.2 von Rapidminer durchgeführt. Im ersten Schritt wird die Datenvorverarbeitung genauer erläutert.

5.2.1 Vorverarbeitung der Daten

Die Datenvorverarbeitung setzt sich aus mehreren Schritten zusammen und orientiert sich an Abschnitt 4.3.1. Dabei muss berücksichtigt werden, dass die Prozesse im entwickelten Vorgehensmodell technisch nicht der Reihenfolge nach genauso umsetzbar sind. Daher kann es Abweichungen in der Reihenfolge geben, inhaltlich werden alle Prozesse aus dem Vorgehensmodell betrachtet. Im ersten Schritt muss ein von Rapidminer lesbares Datum erzeugt werden. In Abbildung 34 ist der Prozess zur Anpassung des Datums zu erkennen.

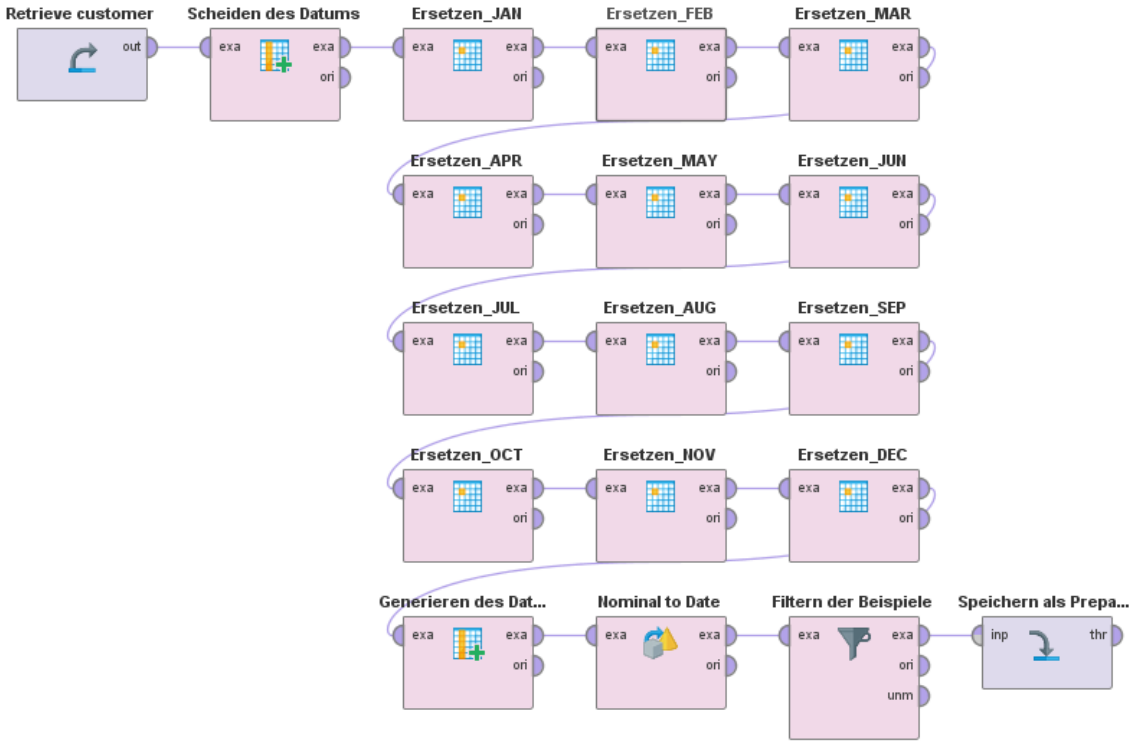


Abbildung 34: Rapidminer: Anpassen des Datum

Der Prozess ist im elektronischen Anhang EA1 zu finden. Als Grundlage wird die Tabelle „customer“ verwendet, mit dem Operator „Retrieve“ wird diese gelesen und den nachfolgenden Operatoren zugeführt. Die Tabelle „customer“ besteht aus über 7.000.000 Zeilen und hat 71 Attribute. Eine Anwendung von Restriktionen findet im ersten Schritt nicht statt, weil ein lesbares Datum zur Anwendung benötigt wird. Im Operator „Schneiden des Datums“ muss das Attribut mit dem

zu untersuchenden Datum (orderDate) in drei Attribute zerlegt werden. Die derzeitige Formatierung des Monats wird von Rapidminer nicht erkannt. Damit entstehen drei neue Attribute mit folgendem Namen: Tag, Monat und Jahr. Für das Attribut des Monats muss jeder vorhandene Wert ersetzt werden. Derzeit findet sich die Bezeichnung beispielsweise für den Monat Januar mit JAN, dies soll ersetzt werden durch 01. Das wird im Operator „Ersetzen_JAN“ umgesetzt und muss nachfolgend für alle zwölf Monate durchgeführt werden. Damit wird aus dem FEB, eine 02 usw. Nach der Formatierung des Monats, können die drei einzelnen Attribute wieder zu einem Attribut DATE zusammengefasst werden. Dafür werden die drei Attribute durch den Operator „Generieren des Datums“ im Format ddMMyy zusammengefügt und im nachfolgenden Operator in den Datentyp Date umgewandelt.
Nachdem das Datum für Rapidminer lesbar gemacht wurde, können die Restriktionen auf die Daten angewendet werden. In der Tabelle 9 sind die angewendeten Restriktionen aufgezeigt.

Tabelle 9: Rapidminer: Anwenden der Restriktionen

Attribut	Vergleichsoperator	Abhängiger Wert
Ordertype	equals	FOR
DATE	≥	06/01/2015
DATE	≠	05/26/2016
storageArea_ID	≠	96789
storageArea_ID	≠	96787
storageArea_ID	≠	96786

Mit der Eingrenzung des Attributs Ordertype werden nur Bestellungen betrachtet, welche vom Kunden ausgelöst wurden. Dadurch werden alle anderen Bestellungen ausgeschlossen, die nicht mit den Kundenbestellungen zusammenhängen. Die nächsten Restriktionen werden auf das Attribut DATE angewendet, dabei werden im ersten Schritt die Daten herausgelöscht, welche nicht repräsentativ sind oder vor dem Betrachtungszeitraum liegen. In diesem Fall werden die Daten ausgeschlossen, weil das Logistikzentrum zu diesem Zeitpunkt in Betrieb genommen wurde. Der zweite Eintrag wird genutzt, weil an diesem Tag die Daten aus dem WMS herausgenommen wurden und die Daten des Tages nicht vollständig verfügbar sind. Die letzten drei Restriktionen beschäftigen sich mit dem Lagerbereich. Dabei gibt die Spalte des abhängigen Wertes, den entsprechenden Lagerbereich in Form einer Nummer aus. Die Lagerbereiche des Lagers für zu große Güter, Gefahrgutlager und spezielle ölige Substanzen werden ausgeschlossen. Nachdem alle Restriktionen angewendet wurden, kann im letzten Schritt die Tabelle unter dem Namen „prepard_v1“ gespeichert werden. Damit kann sie im nachfolgenden Prozess genutzt werden.

In der Abbildung 35 werden die Anzahl der Produkte pro Tag berechnet. Der Prozess befindet sich im elektronischen Anhang EA2.

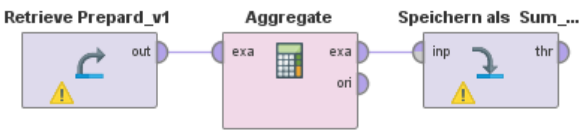


Abbildung 35: Rapidminer: Anzahl der Produkte pro Tag

Als Basis wird die beschriebene Tabelle „prepard_v1“ genutzt, um damit die Berechnungen durchzuführen. Um die Berechnung durchzuführen werden nicht alle Attribute benötigt, der Operator „Aggregate“ betrachtet nur die im Operator ausgewählten Attribute. Dafür werden nun alle verschickten Mengen eines Produktes summiert, welche an einem Tag anfallen. Dieser Wert wird anschließend in Abhängigkeit vom Datum in eine Tabelle geschrieben. Somit entsteht eine Tabelle, bei der für jeden Tag die Anzahl der verschickten Produkte am Tag abgebildet ist. Diese Tabelle wird als „Sum_of_all_Products_per_Day” gespeichert und im nächsten Schritt der Datenvorverarbeitung weiterverwendet.

Der nächste Prozessschritt besitzt einen Unterprozess, in der Abbildung 36 ist der Ausgangsprozess zu erkennen und in Abbildung 37 der Unterprozess. Der gesamte Prozess befindet sich im elektronischen Anhang EA3.

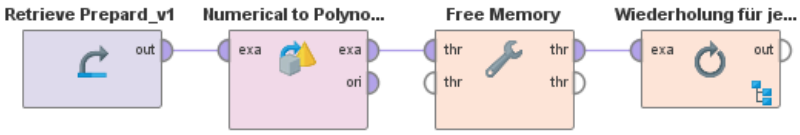


Abbildung 36: Rapidminer: Erzeugen von Tabellen für jede Kategorie

Der Ausgangsprozess in Abbildung 36 startet mit dem Lesen der bereits vorverarbeiteten Tabelle. Darauf folgt eine Transformation des Attributs Kategorie. Das Attribut wird von numerischen Werten zu nominellen Wert transformiert, um das Attribut im Operator „Wiederholung für jede Kategorie“ lesen zu können. Nach der Transformation wird ein Operator eingesetzt, der die nicht genutzten Daten aus dem Arbeitsspeicher entfernt, um eine höhere Rechenleistung zu generieren. Im letzten Operator werden nun die Grundlagen für die Schleife gelegt, welche im Unterprozess dieses Operators durchgeführt wird. Mit Hilfe des Operators wird festgelegt, dass für alle auftretenden Werte der Kategorie der Unterprozess durchgeführt werden soll.

Dieser Unterprozess in Abbildung 37 beginnt mit den Daten, welche in den Schleifenoperator im Ausgangsprozess eingehen. Die Daten haben im Unterprozess einen eigenen Ausgang und stellen den Beginn des Unterprozesses dar. Zur Ermittlung des Zielwertes werden zwei verschiedene Prozesse benötigt, welche sich am Ende in einer Tabelle vereinen. Der erste Prozess beginnt mit den Daten, welche in den Schleifenoperator hineingehen. Zuerst wird die Datenmenge reduziert, indem Attribute ausgeschlossen werden. In diesem Fall werden die Menge, das Datum und die jeweilige Kategorie zur Weiterverarbeitung benötigt.

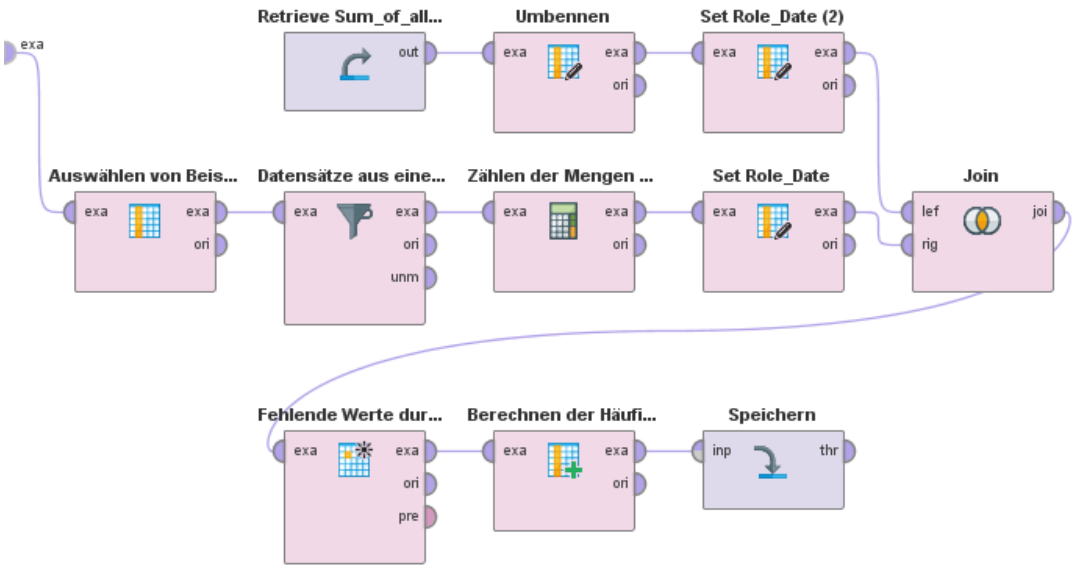


Abbildung 37: Rapidminer: Unterprozess zum Erzeugen von Tabellen für jede Kategorie

Im darauffolgenden Operator werden nur die Datensätze ausgewählt, welche aus einer Kategorie stammen. Mit den ausgewählten Daten soll die Menge der verschickten Produkte in einer Kategorie in Abhängigkeit vom Datum summiert werden. Somit werden alle Produkte betrachtet, welche sich in einer Kategorie befinden und die jeweiligen Mengen der einzelnen Produkte werden summiert. Nach der Aufsummierung werden die Daten in Abhängigkeit vom Datum in eine Tabelle geschrieben. Als Ergebnis steht eine Tabelle mit den aufsummierten Mengen je Kategorie an dem jeweiligen Datum. Neben dem beschriebenen Prozess, wird ein zweiter Prozess im oberen Teil der Abbildung 37 durchgeführt. Dieser Prozess beginnt mit der Tabelle aus Abbildung 35. In dieser Tabelle wird vorerst ein Attribut umbenannt, da sonst die gleichen Attribute beim Zusammenführen der Tabellen vorliegen würden. Bei beiden Prozessen wird das Datum als ID bestimmt um damit mit dem Operator „Join“ beide Tabellen zusammenzuführen.

In diesem Fall wird eine Zusammenführung von der linken (lef am Operator „Join“) Seite aus durchgeführt, weil in dieser Tabelle für jedes Datum ein Eintrag vorliegt. Dies kann bei der anderen Tabelle nicht gewährleistet werden, da nicht jeden Tag ein Produkt aus einer Kategorie verschickt wird. Nachdem beide Tabellen zusammengeführt sind, liegen nun drei Attribute vor: das Datum, die aufsummierten Produktmengen für jeden Tag und die aufsummierten Mengen für eine Kategorie pro Tag. Im nächsten Schritt müssen die fehlenden Werte ersetzt werden. Daher werden alle fehlenden Werte des Attributes von den aufsummierten Mengen für eine Kategorie pro Tag durch null ersetzt. Nachdem keine fehlenden Werte in der Tabelle vorhanden sind, kann die Häufigkeit für jeden Tag der Kategorie berechnet werden. Dafür werden die Mengen je Kategorie durch die summierten Mengen des gesamten Tages geteilt. Als Ergebnis steht die Häufigkeit der Auslagerung für jede Kategorie. Als Ergebnis wird für jede Kategorie die Tabelle abgespeichert, die Tabelle wird bei der Durchführung des NN im nächsten Abschnitt benötigt. Am Ende von diesem Prozess können noch externe Faktoren zu den einzelnen Tabellen hinzugefügt werden.

5.2.2 Zeitreihenprognose mit dem Neuronalen Netz

Für die Durchführung des NN wird ebenfalls der Ausgangsprozess aus Abbildung 36 benötigt, der Ablauf ist identisch zu dem bereits Beschriebenen. Erweitert wird der Prozess lediglich durch zwei Speicheroperatoren um die Leistungsfähigkeit des Modells und das Modell selber zu speichern. Der dargestellte Prozess befindet sich im elektronischen Anhang EA4. Das Ziel besteht darin, dass für jede Kategorie ein Modell des NN generiert wird.

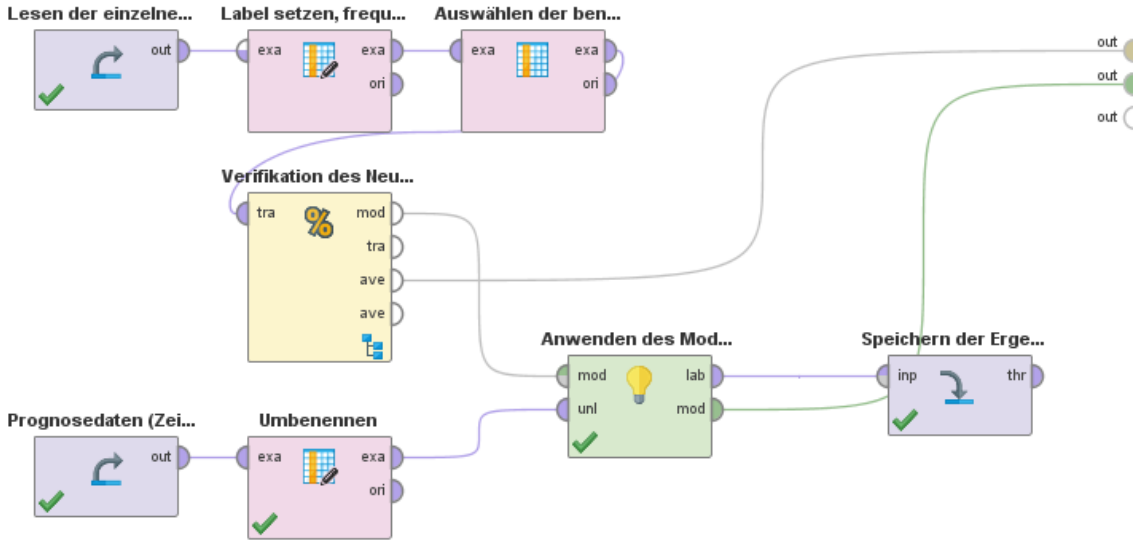


Abbildung 38: Rapidminer: Anwenden und Erzeugen des Neuronalen Netzes

In der Abbildung 38 ist der Prozess in der Schleife zu erkennen, dieser wird für jede Kategorie durchgeführt. In diesem Fall sind wieder zwei verschiedene Prozesse zu erkennen, begonnen wird mit dem oberen Prozess. Dieser beginnt mit dem Lesen der erzeugten Tabellen aus der Datenvorverarbeitung für die in diesem Moment betrachtete Kategorie. Im ersten Operator wird ein Label auf das Attribut der ermittelten Häufigkeit der Kategorie am Tag gesetzt. Darauffolgend werden noch weitere Attribute herausgenommen, da sie nicht notwendig sind für die Erzeugung des Neuronalen Netzes. Diese Tabelle wird nun an den Operator der Verifikation des NN gegeben, dieser enthält ebenfalls wieder einen Unterprozess. Dieser Operator teilt die eingegebenen Datensätze in lokale Trainings-und Testdaten auf, um eine Verifizierung des Modells durchzuführen. Diese Aufteilung erfolgt in diesem Fall zehnmal. Am Ende wird das Modell genommen, welches die beste Leistungsfähigkeit besitzt.

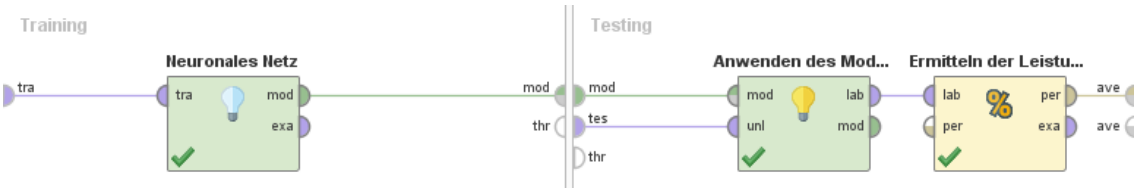


Abbildung 39: Rapidminer: Verifikation des Neuronalen Netz

In Abbildung 39 ist diese Aufteilung zu erkennen, dazu wird mit den lokalen Trainingsdaten zuerst das NN angelernt. Nachdem Lernen des Prozesses wird das Modell auf die Testdaten angewendet. Dabei wird untersucht, wie gut das Modell die Wahrscheinlichkeit vorhersagt. Diese Vorhersage wird mit der Ermittlung der mittleren quadratischen Abweichung gemessen. Je niedriger der Wert, desto besser die Vorhersage. Nach zehnmaligen Durchführen ist der Operator der Verifikation beendet und der Operator liefert die Leistungsfähigkeit und das gelernte Modell.

Um das entwickelte Modell anwenden zu können, wird ein Datensatz benötigt, welcher das Datum oder die Daten des Prognosezeitraumes enthält. Dieser Datensatz wird durch den zweiten Prozess geliefert, dafür wird die Tabelle „Prognosedaten“ gelesen. In der Tabelle existieren zwei Attribute, einmal das zukünftige Datum. Das zweite ist leeres Attribut, welches den gleichen Namen trägt, wie die Häufigkeit je Kategorie aus dem vorhergehenden Prozess. Sollte dies nicht der Fall sein, kann dies über den Operator „Umbenennen“ geändert werden.

Im nächsten Schritt sollen beide Prozesse über den Operator „Anwenden des Modells“ vereinigt werden. In diesem Operator wird das entwickelte NN auf die neuen Daten angewendet. Dadurch entsteht die Prognose. Die daraus erzeugte Tabelle wird unter dem Kategorienamen gespeichert. Dieser beschriebene Prozess muss für jede Kategorie durchgeführt werden. Danach werden die entwickelten Modelle des NN, die Leistungskennwerte und die vorhergesagten Daten gespeichert. Die Daten des Modells des NN befinden sich im elektronischen Anhang EA8. Durch die Menge an Daten und die entsprechenden Abweichungen ist eine Beschreibung an dieser Stelle nicht sinnvoll. Die Durchführung des Vorgehensmodells ist an dieser Stelle beendet, nachfolgend wird jedoch noch die Transformation der Daten zur Produktebene näher erläutert.

5.2.3 Transformation der Ergebnisse

Der Prozess zur Transformation der Ergebnisse auf Produktebene beinhaltet zwei verschiedene Schleifen. Dadurch wird er komplex und schwer nachvollziehbar. Zum besseren Verständnis wird auf die Prozesse im elektronischen Anhang EA5 und EA6 verwiesen. Das Ziel dieser Transformation besteht darin, eine Vorhersage für jedes Produkt treffen zu können.

Um die Transformation erfolgreich durchzuführen werden zwei weitere Tabellen benötigt. Diese Tabellen müssen über einen Prozess erzeugt werden. Dieser vorgelagerte Prozess ist in Abbildung 40 zu erkennen. Ziel ist es eine Tabelle zu erzeugen, die eine Aussage darüber trifft, wie häufig das jeweilige Produkt in einer Kategorie vorkommt.

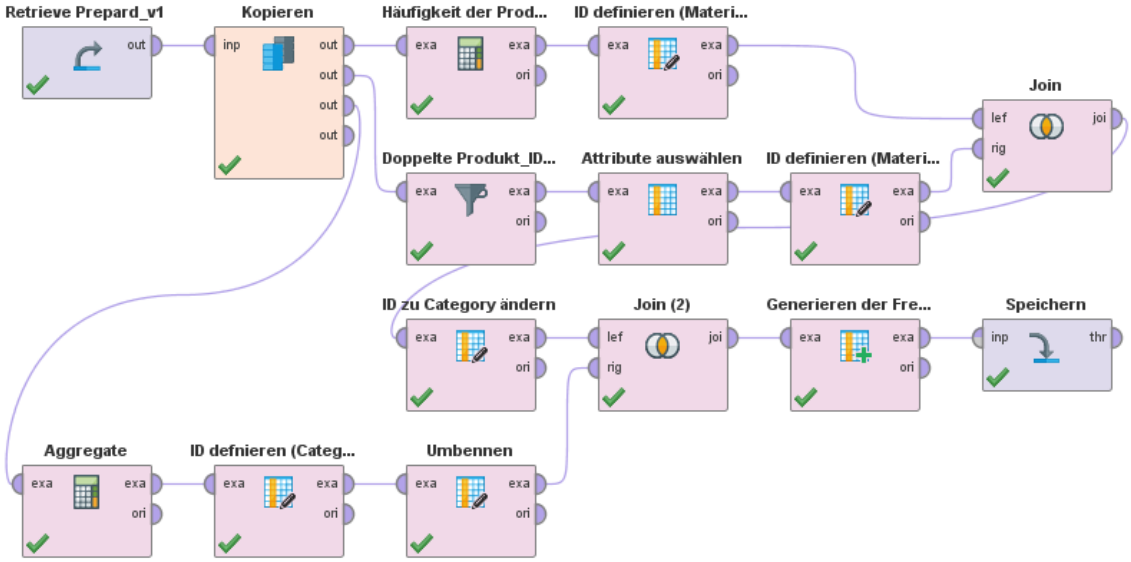


Abbildung 40: Rapidminer: Erzeugen der Tabellen zur Transformation

Dafür wird als Grundlage die Tabelle „prepard_v1“ verwendet, weil sie alle Produkte mit der zugehörigen Kategorie enthält und bereits vorverarbeitet ist. Diese Tabelle wird mit dem Operator „Kopieren“ dreimal kopiert und unterschiedlichen Prozessen zugeführt. In dieser Beschreibung

wird mit dem oberen Prozess (oberste Ausgang am Operator „Kopieren“) begonnen. Mit Hilfe des Operators „Häufigkeit der Produkte“ wird ermittelt, wie häufig das jeweilige Produkt in der Tabelle „prepard_v1“ vorkommt. Letztendlich entstehen zwei Spalten, die eine gibt die Produktnummer (Material_ID) und die zweite gibt die Anzahl des Auftretens der Produktnummer an.

Der zweite Prozess beginnt mit dem zweiten Ausgang aus dem Operator „Kopieren“, dort werden im ersten Schritt die doppelten Werte der Produktnummer entfernt. Daraufhin wird neben der Produktnummer das Attribut Kategorie ausgewählt. Damit ist bekannt, welches Produkt zu welcher Kategorie gehört. Nachfolgend werden der erste Prozess und der zweite Prozess mit Hilfe des Operators „Join“ über die Material_ID zusammengeführt. Daraus entsteht eine Tabelle, welche drei Attribute besitzt: Produktnummer (Material_ID), Kategorie und die Häufigkeit von einem Produkt.

Der dritte Prozess startet ebenfalls mit dem Operator „Kopieren“, dies ist der Dritte und letzte Ausgang. Dabei werden die Produktnummern ebenfalls gezählt, doch diesmal in Abhängigkeit der Kategorie aufgeschrieben. Somit entstehen zwei Attribute, die Kategorie und die Anzahl der Produktnummern in jeder Kategorie. Danach muss ein Attribut umbenannt werden, weil sonst zwei Attribute bei der Zusammenführung den gleichen Namen tragen würden. Die Tabelle aus den ersten beiden Prozessen wird anschließend mit der Tabelle aus dem letzten Prozess über die Material_ID zusammengeführt.

Nach der Zusammenführung wird ein Attribut erzeugt, welches die Auskunft darüber gibt, wie häufig ein Produkt in einer Kategorie vorkommt. Dafür werden die beiden erzeugten Attribute geteilt, die Häufigkeit von einem Produkt durch die Anzahl der Produkte in einer Kategorie. Daraus ergibt sich die Häufigkeit eines Produktes in einer Kategorie. Diese Tabelle wird gespeichert und im nachfolgenden Prozess benötigt.

Diese beiden Tabellen werden im Prozess benötigt, welcher die Wahrscheinlichkeit zur Auslagerung für jedes Produkt angibt. Der Prozess enthält zwei Schleifen, dafür werden die Werte für die erste Schleife identisch zu Abbildung 36 erzeugt. Die erste Schleife wird für jede Kategorie durchgeführt und die zweite für jedes Produkt in dieser Kategorie. Dafür wird der Prozess in Abbildung 41 benötigt, begonnen wird mit der erzeugten Tabelle aus Abbildung 40.

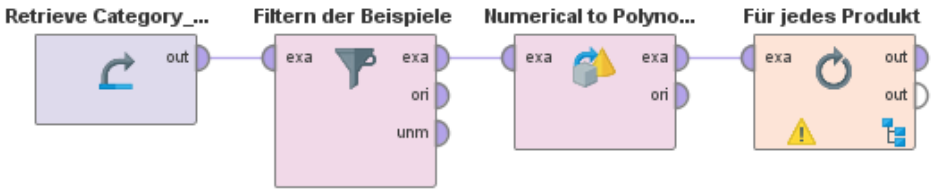


Abbildung 41: Rapidminer: Erzeugen einer Schleife für jedes Produkt

Aus dieser Tabelle werden alle Datensätze herausgefiltert, die der Kategorie zugehörig sind, welche sich in der Schleife angeschaut wird. Somit werden nur die Produkte aus einer Kategorie betrachtet. Daraufhin wird eine weitere Schleife entwickelt, welche für jedes Produkt in der ausgewählten Kategorie durchgeführt wird.

Die Abbildung 42 hat ebenfalls zwei Prozesse, welche getrennt voneinander betrachtet werden müssen. Begonnen wird mit dem oberen Prozess, dabei werden die erzeugten Tabellen aus dem NN gelesen. In diese Tabelle wird nun ein leeres Attribute mit dem Namen Kategorie eingefügt. Im darauffolgenden Schritt wird in diese Tabelle die derzeit betrachtete Kategorie geschrieben, weil die Kategorie die ID zum späteren Zusammenführen darstellt.

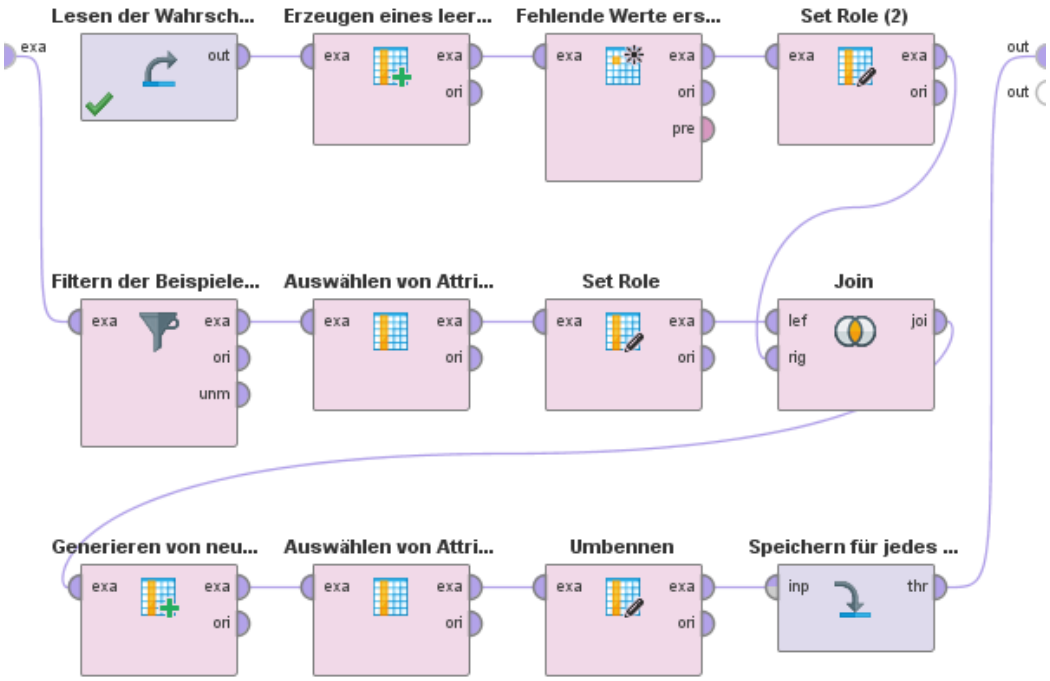


Abbildung 42: Rapidminer: Ermittlung der Wahrscheinlichkeit der Auslagerung für jedes Produkt

Der zweite Prozess bekommt seine Tabelle aus dem übergeordneten Prozess, dabei wird im ersten Schritt das gerade in der Schleife betrachtete Produkt herausgefiltert. Aus dieser gefilterten Tabelle werden nun zwei Attribute ausgewählt, die Kategorie und die Häufigkeit des Produktes in der Kategorie. In beiden Prozessen ist die Kategorie die ID und über diese werden beide Tabellen zusammengeführt. Da jedoch aus dem zweiten Prozess nur eine Zeile vorliegt, muss diese zu jeder Zeile der zweiten Tabelle hinzugefügt werden. Deshalb findet in dem Operator „Join“ eine Zusammenführung von dem ersten Prozess aus statt. Dadurch entsteht eine Tabelle mit vier Attributen, dem Datum, der Kategorie, der Wahrscheinlichkeit der Auslagerung und der Häufigkeit des Produktes in der Kategorie. Nun wird aus den beiden zuletzt genannten Attributen durch Multiplikation die Wahrscheinlichkeit der Auslagerung für jedes Produkt bestimmt. Danach werden alle Attribute entfernt, bis auf das Datum und die Wahrscheinlichkeit für die Auslagerung. Das Attribut, indem die Wahrscheinlichkeit für die Auslagerung steht wird umbenannt in die Produktnummer und zum Schluss unter dem Namen der Produktnummer gespeichert.

Durch den beschriebenen Prozess entsteht für jedes Produkt eine eigene Tabelle, das entspricht in dieser prototypischen Umsetzung 127.314 Tabellen. Die Aufgabe des DM ist in diesem zu diesem Zeitpunkt abgeschlossen, die Nutzfreundlichkeit ist jedoch nicht vorhanden. Daher wird im Folgenden eine Möglichkeit gesucht, diese Tabellen zu einer zusammenzuführen. Dadurch entsteht eine Tabelle mit 127.315 Attribute. Für jedes Produkt wird ein eigenes Attribut erzeugt und ein Attribut, welches das vorherzusagende Datum wiedergibt. In Abbildung 44 ist die Zusammenführung der Daten dargestellt. Die prototypische Umsetzung des Prozesses findet sich in Abbildung 44 wieder und der Prozess selber im elektronischen Anhang EA7.

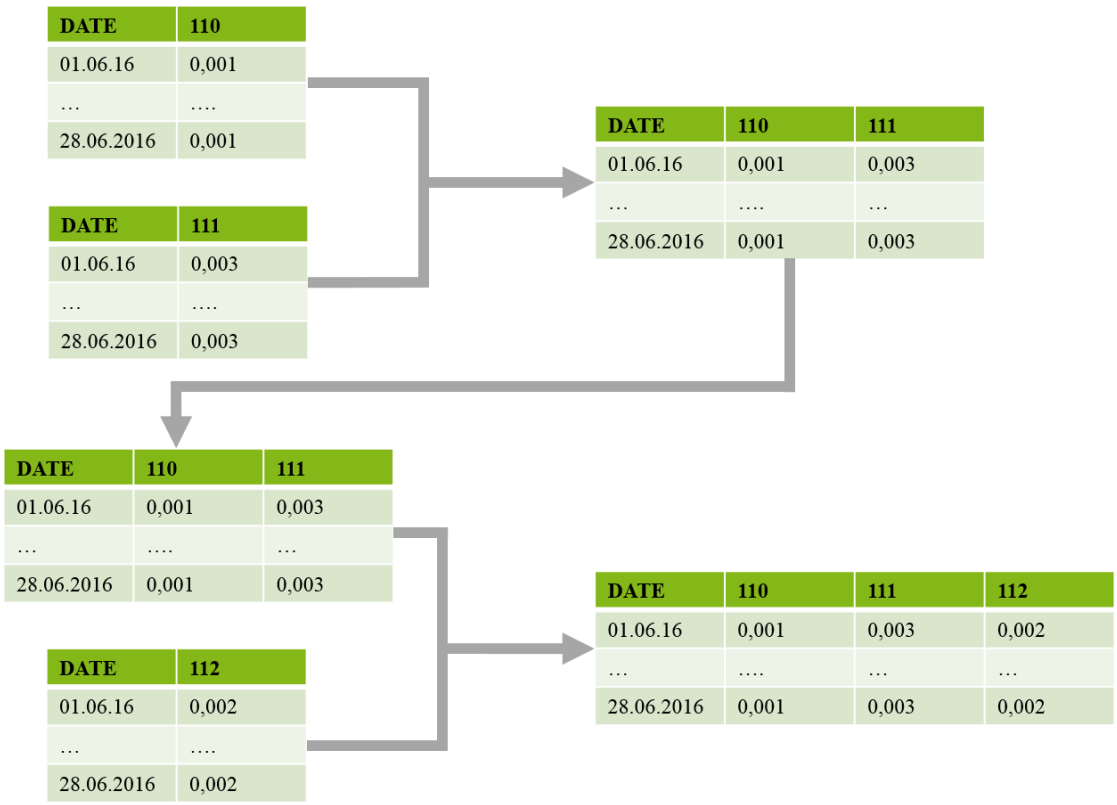


Abbildung 43: Verfahren zur Vereinigung der Tabellen

Derzeit liegen über 100.000 Tabellen mit jeweils zwei Spalten vor. Im ersten Schritt müssen zwei Tabellen über das Attribut „DATE“ vereinigt werden. Dadurch entsteht eine Tabelle mit drei Attributen, ein Attribut gibt das Datum wieder, die anderen beiden jeweils die Wahrscheinlichkeit der Auslagerung des Produktes. Dadurch wurden bereits zwei Produkte mit ihrer Wahrscheinlichkeit zur Auslagerung in einer Tabelle zusammengefügt. Die erzeugte Tabelle wird im nachfolgenden Prozess benötigt, an die Tabelle mit den drei Attributen soll eine weitere Tabelle mit zwei Attributen hinzugefügt werden. Bei der Zusammenführung der beiden Tabellen entsteht daraus wieder eine Tabelle mit vier Attributen. Dabei ist ein weiteres Attribut für die Wahrscheinlichkeit der Auslagerung eines Produktes hinzugekommen. Der beschriebene Prozess muss für alle vorhandenen Tabellen durchgeführt werden, bis alle vorhandenen Tabellen in einer Tabelle vereinigt wurden. Auf Basis der beschriebenen Idee, wird ein Rapidminer Prozess entwickelt.

Dieser entwickelte Prozess ist in Abbildung 44 verdeutlicht. Begonnen wird bei diesem Prozess mit den Daten als iOObject (Collection of Collection of Data Tables). In einem iOObject sind die Daten als Array gespeichert. In diesem ersten Operator sind somit alle Produkte mit ihren zugehörigen prognostizierten Werten vorhanden. Diese müssen im ersten Schritt geglättet werden, weil derzeit noch die Tabellen der Produkte in Abhängigkeit der Kategorie gespeichert sind. Der nachfolgende Algorithmus benötigt die Tabellen alle direkt in der ersten Ebene (Collection of Data Tables). Deshalb wird die Ebene mit den Kategorien entfernt und alle Tabellen der Produkte werden direkt verfügbar für den nächsten Operator gemacht. Dort werden die Daten multipliziert und dann an zwei Ausgänge abgegeben. Die Herausforderung besteht darin, die erste Vereinigung von den Tabellen durchzuführen. Um dies zu erreichen, muss der Datensatz multipliziert werden, damit im oberen Prozess der erste Datensatz aus dem Array herausgesucht werden kann. Dieser wird mit Hilfe des nachfolgenden Operators zwischengespeichert und im späteren Verlauf

noch einmal aufgegriffen. Der Operator zum Zwischenspeichern besitzt einen Namen unter dem gespeichert wird, dies entspricht in diesem Fall dem Zahlenwert 1.

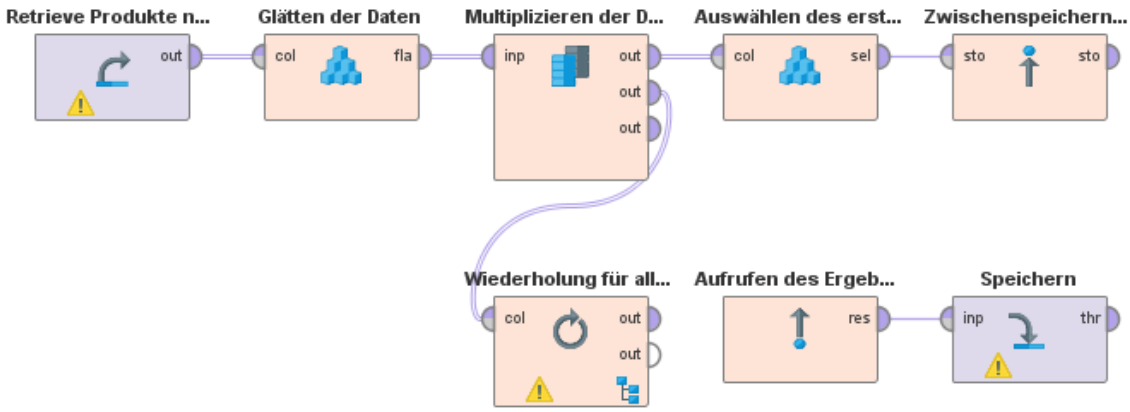


Abbildung 44: Rapidminer: Zusammenführen einzelner Tabellen zu einer gesamten Tabelle

Da alle Tabellen betrachtet werden sollen, muss der Vereinigungsprozess für alle Tabellen durchgeführt werden. Dadurch wird ein Operator benötigt, welcher eine Schleife abbildet und mit einem Array umgehen kann. Der ausgewählte Operator (Wiederholung für alle Tabellen) enthält einen Unterprozess. Mit der Einführung des „Wiederholung für alle Tabellen“ Operators wird ebenfalls ein Makro gesetzt, welches automatisch von eins beginnend hochzählt. Dies wird benötigt um den in Abbildung 43 beschriebenen Prozess durchführen zu können. Die Umsetzung in Rapidminer wird in Abbildung 45 genauer erläutert. Der Unterprozess besitzt nur einen Operator, welcher dazu dient eine Wenn-Dann Situation abzubilden. Ausgehend von einer Bedingung wird entweder Prozess 1 (Then) oder Prozess 2 (Else) ausgeführt, dies ist in Abbildung 45 verdeutlicht.

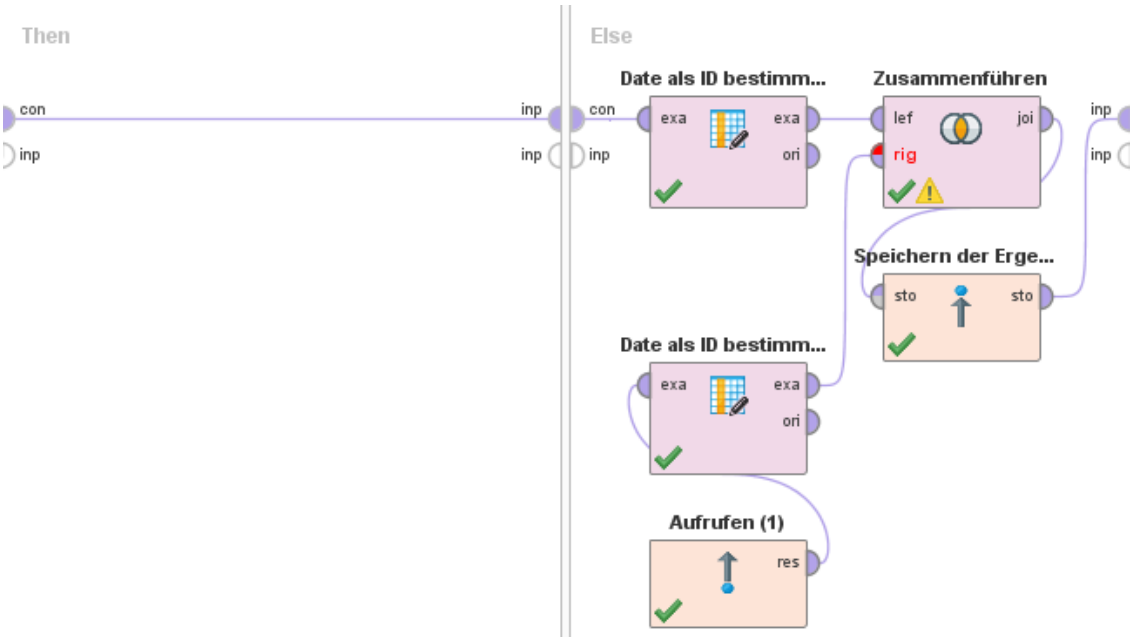


Abbildung 45: Rapidminer: Auswahl des Prozess 1 (Then) oder Prozess 2 (Else)

Zuerst muss die Wenn-Bedingung definiert werden, nach welcher der Prozess ausgewählt werden soll. Die Bedingung lautet, sofern das definierte Makro eins ist, dann wird der Prozess 1 genommen, ist er abweichend von der Zahl eins, wird der Prozess 2 genommen. Damit wird ausgeschlossen, dass die in Abbildung 43 im oberen Prozess ausgewählte Tabelle ein zweites Mal genutzt wird, weil der erste Datensatz keinen Prozess zugeführt wird und direkt an die Schleife

zurückgegeben wird. Daher erhöht das Makro den Wert auf 2 und gleichzeitig wird der zweite Datensatz aus dem Array genommen. Dabei wird als erstes im unteren Teil der Abbildung 45 die selektierte Tabelle aus Abbildung 44 aufgerufen. Dabei ist zu beachten, dass die Operatoren zum zwischenspeichern und aufrufen jeweils über einen Namen verbunden sind, welcher in dieser Situation dem Zahlenwert 1 entspricht (vgl. Abbildung 44 oberer Prozess). Für die aufgerufene Tabelle wird als ID das Datum definiert. Das Makro hat bereits zur Zahl 2 heraufgezählt. Daher ist die Bedingung ungültig und es wird Prozess 2 durchgeführt. Aus diesem Grund wird ebenfalls der zweite Datensatz aus dem Array angeschaut und mit einer ID für das Datum versehen. Nun werden beide Tabellen miteinander vereinigt und abschließend wird die erzeugte Tabelle über einen Operatoren zum zwischenspeichern festgehalten. Der Name wird über das Makro bestimmt und entspricht in diesem Fall dem Zahlenwert 2. Verglichen mit der Abbildung 43 wurde eine Tabelle mit 3 Attributen entwickelt und zwischengespeichert. Nach Beendigung des Prozesses zählt das Makro eine Zahl höher und der Prozess beginnt von vorne mit einer Ausnahme. Im Operator „Aufrufen“ wurde ebenfalls der Name mit der Beendigung des vorhergehenden Prozess auf zwei angepasst, dadurch wird nun die Tabelle mit den drei Attributen aufgerufen. Aus dem Array wird der dritte Datensatz hinzugezogen und mit der Tabelle mit den drei Attributen vereinigt. Dieser Prozess wird für alle Tabellen durchgeführt. Sofern für alle Elemente des Arrays der Prozess durchgeführt wurde, wird die ermittelte und zusammengeführte Tabelle gespeichert und der Prozess endet von selbst.

Insbesondere der letzte Teil stellt einen komplexen Prozess dar, deshalb wird insbesondere nochmal auf die parallele Nutzung von Rapidminer zu dieser Arbeit hingewiesen.

6 Zusammenfassung und Ausblick

Die Grundlagen zur Optimierung des Einlagerungsproblems wurden in den ersten beiden Kapiteln gelegt. Zu Beginn wurden die Grundlagen der Einlagerung erläutert. Dafür wurden die Prozesse in einem teilautomatisierten Logistikzentrum beschrieben. Hierbei wurde auf spezielle Einlagerungsstrategien eingegangen und es wurde untersucht, wie das DM in den beschriebenen Algorithmen eingesetzt werden kann. Parallel wurde ein Vorgehensmodell vorgestellt, welches auf Basis von DM die Einlagerung bestimmt. Ebenfalls wurde im ersten Literaturkapitel das zugrundeliegende Entscheidungsproblem eingeordnet und näher erläutert. In dieser Arbeit wurden verschiedene Vorgehensmodelle zur Anwendung von DM entwickelt, dafür wurden die Grundlagen im dritten Kapitel gelegt. Dabei wurde ein KDD-Vorgehensmodell nach [FPS96] vorgestellt. Ebenfalls wurden detaillierte Maßnahmen zur Vorverarbeitung von Daten für die DM-Verfahren aufgezeigt. Die DM-Verfahren des NN, der SVM und des Entscheidungsbaumes wurden näher erläutert, um sie erfolgreich in dem nachfolgenden Kapitel anwenden zu können. Beendet wird das dritte Kapitel mit Kennwerten zur Messung der Leistungsfähigkeit der einzelnen Verfahren.

Auf Basis der beiden Möglichkeiten zur Bestimmung des Produktlagerortes wurde eine spezifische Problemstellung definiert. Aus dieser Problemstellung heraus wurden zwei Handlungsalternativen zur Optimierung des Einlagerungsproblems entwickelt. Dabei stellen beide Handlungsalternativen ein Vorgehensmodell zur Anwendung der DM-Verfahren dar, welche sich auf das Vorgehensmodell von [FPS96] beziehen. Das Domänenverständnis ist für beide Handlungsalternativen gleich und zeigt die Prozesse und die daraus resultierenden Restriktionen in einem teilautomatisierten Logistikzentrum auf. Die erste Handlungsalternative ermittelt den Lagerbereich durch die Anwendung von DM-Verfahren. Nach der Datenvorverarbeitung wurden für die Daten drei DM-Modelle entwickelt. Auf Basis der Leistungsfähigkeit wurde eine Entscheidung zur Anwendung des NN getroffen. Bei dieser Handlungsalternative ist hervorzuheben, dass Ursprungsdaten benötigt werden, welche den korrekten Lagerbereich angeben. Diese wurden in den vorhergehenden Abschnitten als globale Trainingsdaten bezeichnet. Dabei stellt die Notwendigkeit nach diesen Daten die größte Schwachstelle neben der Intransparenz des Verfahrens dar. Solche Daten werden bei der zweiten Handlungsalternative nicht benötigt, da als Grundlage Vergangenheitsdaten genutzt werden. Die Vergangenheitsdaten geben Auskunft über die Häufigkeit der Auslagerung der Produkte. Daher ist das Ergebnis dieser Handlungsalternative ein Wahrscheinlichkeitswert für die Auslagerung des Produktes. Dieser wird durch die Datenvorverarbeitung und das Anwenden eines NN für jedes Produkt ermittelt. Die Herausforderung in dieser Handlungsalternative liegt in der hohen Anzahl an Produkten in einem teilautomatisierten Logistikzentrum. Für jedes Produkt muss eine eigene Tabelle mit der zukünftigen Wahrscheinlichkeit der Auslagerung angelegt werden. Ebenfalls muss bei diesem Verfahren der Wert in einen bestehenden oder neu generierten Algorithmus eingefügt werden, welcher den Lagerbereich bestimmt. Im Vergleich der beiden Handlungsalternativen wurden bei der ersten Handlungsalterna-

tive fehlende Transparenz und eine fehlende einfache Validierungsmöglichkeit festgestellt. Aufgrund dieser Mängel und der Schwierigkeit geeignete Daten zu bekommen kam es in dieser Arbeit zu einer Entscheidung für die zweite Handlungsalternative.

Diese Entscheidung stellt die Basis für die prototypische Umsetzung im letzten Kapitel dar. Das entwickelte Vorgehensmodell aus Handlungsalternative Zwei wird mit Hilfe von Rapidminer durchgeführt. Die daraus resultierenden Ergebnisse werden in einen Algorithmus zur Bestimmung des Lagerbereiches implementiert.

Somit wurde ein allgemeingültiges Vorgehensmodell zur Verbesserung des Einlagerungsprozesses definiert. Die Durchführbarkeit ist dabei durch die Anwendung der prototypischen Umsetzung aufgezeigt. Eine Implementierung in einen bestehenden Algorithmus muss unabhängig von dieser Arbeit durchgeführt werden. Dabei ist die Implementierung in den jeweiligen Algorithmus abhängig von dem betrachteten teilautomatisierten Logistikzentrum. Das Thema der Lagerplatzoptimierung oder diese Arbeit betreffend, die Zuordnung eines Lagerbereiches sind dauerhaft zu optimierende Prozesse und lassen sich nur teilweise standardisieren. Jedes Logistikzentrum besitzt eigene Restriktionen und Besonderheiten, welche bei der Optimierung des Einlagerungsprozesses für jedes Logistikzentrum separat betrachtet werden müssen. Aus diesen Gründen ist die Anwendung des entwickelten Vorgehensmodells vorher genau zu prüfen.

Durch das Standardisierungsproblem wird der Ausblick auf zwei verschiedene Arten präsentiert. Ein Teil des Ausblickes zeigt, wie die in dieser Arbeit beschriebenen Vorgehensmodelle noch verändert werden können. Der zweite Teil gibt Auskunft über mögliche weitere Forschungstätigkeiten gegenüber des Themas der Einlagerungsoptimierung.

Das erste entwickelte Vorgehensmodell hat Schwachstellen bei den Validierungsmöglichkeiten, deshalb muss in diesem Bereich weiter geforscht werden. Dabei soll der iterative Prozess zwischen der Simulation und dem DM betrachtet werden. Daraus kann ein erweitertes Vorgehensmodell entstehen, welches ohne manuelle Eingriffe die Simulation mit dem DM verbindet. Ebenfalls ist bei beiden Handlungsalternativen denkbar, dass die Ergebnisse einer Warenkorbanalyse mit in die Ergebnisse der Handlungsalternativen einfließen. Die Warenkorbanalyse untersucht, wie häufig Produkte zusammen gekauft werden. Das Ziel besteht darin, eine genauere Vorhersage treffen zu können. Ein möglicher Ansatz hierzu wäre, die Handlungsalternative Zwei durchzuführen und ein weiteres Vorgehensmodell zur Anwendung der Warenkorbanalyse zu entwickeln. Ein Zusammenfassen von beiden Vorgehensmodellen stellt das Ergebnis in Form einer Wahrscheinlichkeit für die Auslagerung dar. Dabei sollte der Schwerpunkt auf die Verknüpfung der beiden Vorgehensmodelle gelegt werden.

In dieser Arbeit wurde mit überwachten Lernverfahren gearbeitet. Weiterhin besteht die Forschungsmöglichkeit zur Verknüpfung von Verfahren des unüberwachten und überwachten Lernens. Das Ziel der unüberwachten Lernverfahren besteht darin, globale Trainingsdaten für die erste Handlungsalternative zu finden. Die unüberwachten Lernverfahren benötigen keine Trainingsdaten um ein Modell mit entsprechenden Ergebnissen zu erzeugen. Dadurch können die globalen Trainingsdaten durch das unüberwachte Lernverfahren erzeugt werden und in den Verfahren des überwachten Lernens angewendet werden.

Den letzten Teil stellt die kritische Würdigung dieser Arbeit dar, hierbei wird die Leistung des Autors kritisch betrachtet und es findet eine Einordnung des wissenschaftlichen Mehrwertes statt. Der Arbeitsaufwand dieser Arbeit befasst sich mit der Literaturrecherche, der Entwicklung

von zwei Handlungsalternativen und der prototypischen Umsetzung einer der Handlungsalternativen. Für die zweite Handlungsalternative konnte eine Optimierung des Einlagerungsprozesses festgestellt werden, welches durch ein Validierungsverfahren bestätigt wurde. Die Implementierung und die erfolgreiche Anwendung in einen Algorithmus waren in dieser Arbeit nicht gefordert. Eine Anwendung des entwickelten Vorgehensmodells auf weitere teilautomatisierte Logistikzentren ist in dieser Arbeit nicht betrachtet worden, daher ist eine theoretische Allgemeingültigkeit dieses Vorgehensmodells festzuhalten. Eine Betrachtung des Einlagerungsproblems in Zusammenhang mit dem DM wurde in dieser Arbeit zum ersten Mal durchgeführt. Dadurch ist der wissenschaftliche Anspruch an diese Arbeit erfüllt, da Möglichkeiten zur Optimierung der Entscheidung zur Einlagerung aufgezeigt werden.

Literaturverzeichnis

[Aue16] Auer, B.: Zeitreihe (Version 9). http://wirtschaftslexikon.gabler.de/Archiv/57589/zeitreihe-v9.html, 31.08.2016.

[BC06] Beekmann, F.; Chamoni, P.: Verfahren des Data Mining. In: Chamoni, P.; Gluchowski, P. (Hrsg.): Analytische Informationssysteme. Springer-Verlag GmbH, Berlin, 2006; S. 263–282.

[BK14] Beierle, C.; Kern-Isberner, G.: Methoden wissensbasierter Systeme. Grundlagen, Algorithmen, Anwendungen. Springer Fachmedien, Wiesbaden, 2014.

[BZG+16] Benuwa, B. B. et al.: A Review of Deep Machine Learning. In: International Journal of Engineering Research in Africa, 2016, Vol. 24; S. 124–136.

[Cao10] Cao, L.: Domain driven data mining. Springer-Verlag, New York, 2010.

[CL14] Cleve, J.; Lämmel, U.: Data Mining. De Gruyter Oldenbourg, München, 2014.

[Cro10] Crone, S. F.: Neuronale Netze zur Prognose und Disposition im Handel. Gabler Verlag / GWV Fachverlage GmbH, Wiesbaden, 2010.

[CV95] Cortes, C.; Vapnik, V.: Support-vector networks. In: Machine Learning, 1995, Vol. 20; S. 273–297.

[DEL14] Deuse, J.; Erohin, O.; Lieber, D.: Wissensentdeckung in vernetzten, industriellen Datenbeständen. In: Lödding H. (Hrsg.): Industrie 4.0. Wie intelligente Vernetzung und kognitive Systeme unsere Arbeit verändern. Gito, Berlin, 2014; S. 373–395.

[FPM92] Frawley W. J.; Piatetsky-Shapiro G.; Matheus C. J.: Knowledge Discoveryin Databases: An Overview. In: AI Magazine, 1992, Vol. 13; S. 57–60.

[FPS96] Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.: From Data Mining to Knowledge Discovery in Databases. In: Communications of the ACM, 1996, Vol. 39; S. 37–54.

[FS89] Frazelle E.; Sharp G.: Correlated assignment strategy can improve orderpicking operation. In: Industrial Engineering, 1989, Vol. 21; S. 33–37.

[Gar05] Garfinkel, M.: Minimizing Multi-zone Orders in the Correlated Storage Assignment Problem. Georgia Institute of Technology, 2005.

[GK13] Grünig, R.; Kühn, R.: Entscheidungsverfahren für komplexe Probleme. Ein heuristischer Ansatz. Springer Gabler, Berlin, 2013.

[GLH15] García, S.; Luengo, J.; Herrera, F.: Data Preprocessing in Data Mining. Springer International Publishing, Cham, 2015.

[Gra80] Granger, C.: Long memory relationships and the aggregation of dynamic models. In: Journal of Econometrics, 1980, Vol. 14; S. 227–238.

[Hes63] Heskett, J. L.: Cube-Per-Order Index — A Key To Warehouse Stock Location. In: Transportation and Distribution Management, 1963, Vol. 3; S. 27–31.

[HH13] Heinemann, G.; Haug, K. Hrsg.: Digitalisierung des Handels mit ePace: Innovative E-Commerce-Geschäftsmodelle und digitale Zeitvorteile. Springer Fachmedien, Wiesbaden, 2013.

[HM82] Hanley, J. A.; McNeil, B. J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. In: Radiology, 1982, Vol. 143; S. 29–36.

[HS08] Hompel, M.; Schmidt, T.: Warehouse Management. Organisation und Steuerung von Lager- und Kommissioniersystemen. Springer-Verlag, Berlin, 2008.

[KBW+10] Kofler, M. et al.: Reassigning Storage Locations in a Warehouse to Optimize the Order Picking Process. In: European Modeling and Simulation Symposium, 2010, Vol. 22.

[KBW+14] Kofler, M. et al.: Affinity Based Slotting in Warehouses with Dynamic Order Patterns. In: Klempous, R. et al. (Hrsg.): Advanced Methods and Applications in Computational Intelligence. Springer International Publishing, Heidelberg, 2014; S. 123–143.

[KL76] Kallina, C.; Lynn, J.: Application of the Cube-per-Order Index Rule for Stock Location in a Distribution Warehouse. In: Interfaces, 1976, Vol. 7; S. 37–46.

[KS08] Kim B.; Smith J : Dynamic slotting for zone-based distribution center picking operation. In: International Material Handling Research Colloquium, 2008, Vol. 10; S. 577–599.

[Kuh95] Kuhn, A.: Prozeßketten in der Logistik. Entwicklungstrends und Umsetzungsstrategien. Verl. Praxiswissen, Dortmund, 1995.

[LGS14] Laux, H.; Gillenkirch, R. M.; Schenk-Mathes, H. Y.: Entscheidungstheorie. Springer-Verlag, Berlin, Heidelberg, 2014.

[Mar14] Martin, H.: Transport- und Lagerlogistik. Planung, Struktur, Steuerung und Kosten von Systemen der Intralogistik. Springer Vieweg, Wiesbaden, 2014.

[Mar16] Markgraf, D.: Programmbreite. http://wirtschaftslexikon.gabler.de/Archiv/124962/programmbreite-v4.html, 24.08.2016.

[ML13] Müller, R. M.; Lenz, H.-J.: Business intelligence. Springer Vieweg, Berlin u.a., 2013.

[MP43] McCulloch, W. S.; Pitts, W.: A logical calculus of the ideas immanent in nervous activity. In: The Bulletin of Mathematical Biophysics, 1943, Vol. 5; S. 115–133.

[MR10] Maimon, O.; Rokach, L.: Data mining and knowledge discovery handbook. Springer, New York, 2010.

[MSH07] Mantel, R. J.; Schuur, P. C.; Heragu, S. S.: Order oriented slotting. A new assignment strategy for warehouses. In: European Journal of Industrial Engineering, 2007, Vol. 1; S. 301–315.

[Pet09] Petersohn, H.: Data Mining. Verfahren, Prozesse, Anwendungsarchitektur. Oldenbourg, München, 2009.

[PH04] Pil, F. K.; Holweg, M.: Linking Product Variety to Order-Fulfillment Strategies. In: Interfaces, 2004, Vol. 34; S. 394–403.

[Pia10] Piazza, F.: Data Mining im Personalmanagement. Eine Analyse des Einsatzpotenzials zur Entscheidungsunterstützung. GWV Fachverlage GmbH, Wiesbaden, 2010.

[Qui86] Quinlan, J. R.: Induction of Decision Trees. In: Machine Learning, 1986, Vol. 1; S. 81–106.

[RCB14] Rushton, A.; Croucher, P.; Baker, P.: The Handbook of Logistics and Distribution Management. Understanding the Supply Chain. Kogan Page, London, 2014.

[RHW86] Rumelhart, D. E.; Hinton, G. E.; Williams, R. J.: Learning representations by back-propagating errors. In: Nature, 1986, Vol. 323; S. 533–536.

[RN16] Ramesh J.; Narayanan P.: Intelligent Slotting for the Warehouse. In: Kamath, N.; Saurav, S. (Hrsg.): Handbook of research on strategic supply chain management in the retail industry. Business Science Reference, Hershey, PA, 2016; S. 326–342.

[Ros58] Rosenblatt, F.: The perceptron. A probabilistic model for information storage and organization in the brain. In: Psychological Review, 1958, Vol. 65; S. 386–408.

[Run15] Runkler, T.: Data mining. Modelle und Algorithmen intelligenter Datenanalyse. Springer Vieweg, Wiesbaden, 2015.

[Sch94] Schröder, M.: Einführung in die kurzfristige Zeitreihenprognose und Vergleich der einzelnen Verfahren. In: Mertens, P. (Hrsg.): Prognoserechnung. Physica-Verlag, Heidelberg, 1994; S. 7–39.

[SGS14] Schmid, U.; Görz, G.; Schneeberger, J.: Handbuch der Künstlichen Intelligenz. Oldenbourg, München, 2014.

[Sha13] Sharafi, A.: Knowledge Discovery in Databases. Eine Analyse des Änderungsmanagements in der Produktentwicklung. Springer Fachmedien, Wiesbaden, 2013.

[SS04] Smola, A. J.; Schölkopf, B.: A tutorial on support vector regression. In: Statistics and Computing, 2004, Vol. 14; S. 199–222.

[Til03] Tillmanns, C.: Data Mining zur Unterstützung betrieblicher Entscheidungsprozesse. Universität Dortmund, 2003.

[VKS12] Vahrenkamp, R.; Kotzab, H.; Siepermann, C.: Logistik. Management und Strategien. Oldenbourg, München, 2012.

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AUC	Area under the Curve
DM	Data Mining
ER-Modell	Entity-Relationship-Modell
FPR	Falsch-positiv-Rate
KDD	Knowledge in Discovery Databases
k-nn	Nächster-Nachbar-Klassifikator
NN	Neuronales Netz
OOS	order oriented slotting
ROC	Receiver Operating Characteristic
SVM	Support-Vector-Machine
TPR	Richtig-positiv-Rate
WMS	Warehouse Management System

Anhang

Der Anhang dieser Arbeit unterliegt einem Sperrvermerk und liegt in einem zweiten Dokument vor.

Eidesstattliche Versicherung

Name, Vorname | Matr.-Nr.

Ich versichere hiermit an Eides statt, dass ich die vorliegende Masterarbeit mit dem Titel:

Optimierung des Entscheidungsprozesses zur Einlagerung von Produkten in einem teilautomatisierten Logistikzentrum unter Anwendung von Data Mining

selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Ort, Datum | Unterschrift

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden. (§ 63 Abs. 5 Hochschulgesetz - HG -)

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird gfls. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin“) zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Ort, Datum | Unterschrift