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

Lepaskar Sivalingam, B.Sc.

Studiengang: Wirtschaftsingenieurwesen

Matrikel-Nr.: 137500

Optimierung der Bauteilqualität anhand von Prozessdatenauswertung unter Anwendung der Data-Mining-Verfahren

ausgegeben am: 11.08.2017

eingereicht am: 06.02.2018

Erstprüfer: Prof. Dr.-Ing Markus Rabe

Zweitprüfer: Dipl.-Inf. Reza Jalali

Inhaltsverzeichnis

Abkürzungsverzeichnis

AOI	Automatische optische Inspektion
AV	Average within cluster distance
BDE	Betriebsdatenerfassung
DBI	Davies Bouldin Index
DM	Data-Mining
DMM	Data-Mining-Model
FIFO	First In-First Out
IuK	Informations- und Kommunikationstechnik
KDD	Knowledge Discovery in Database
KDID	Knowledge Discovery in Industrial Database
MES	Manufacturing Execution System
ppm	parts per million (Gebrauch im QM und bei Gase)
QM	Qualitätsmanagement
OGW	Obere Toleranzgrenze
PSN	Product Serial Number
SMT	Surface mounted technology
SPC	Statistische Prozesskontrolle
THT	Through-hole technology
UGW	Untere Toleranzgrenze

Abbildungsverzeichnis

Tabellenverzeichnis

1.Einleitung

Die industrielle Produktion hat für den Wirtschaftsstandort Deutschland eine hohe Bedeutung und ist für seine Herkunftsbezeichnung „Made in Germany“- welche für Qualität und Zuverlässigkeit steht - bekannt. Heutzutage müssen sich die Industrieunternehmen mit zwei Trends auseinandersetzen: Globalisierung und Dynamisierung der Produktlebenszyklen. Durch die zunehmende Globalisierung befinden sich produzierende Unternehmen im Verdrängungswettbewerb und unter stetigem Kostendruck. Die Verkürzung der Zeiträume der Produktlebenszyklen - zwischen zwei Produktgenerationen- führt zu einer geringeren Stückzahl und zeitgleichen Erhöhung der Variantenvielfalt pro Produkt. Das Ziel ist profitabel zu wachsen und gleichzeitig die Position gegenüber den direkten Konkurrenten langfristig zu sichern. Um weiterhin konkurrenzfähig zu bleiben und dem enormen Wettbewerbs- und Preisdruck, besonders aus den Schwellen- und Entwicklungsländern mit ihren niedrigen Löhnen und Produktionskosten, entgegenzuwirken, müssen die internen Kosten niedrig gehalten und die Produktangebote an den Markt angepasst werden. Wichtige Erfolgsfaktoren für die Optimierung interner Kosten sind transparente und kontinuierlich verbesserte Fertigungsprozesse. Bestehende Ansätze wie Lean Management decken jedoch nicht alle Verbesserungspotenziale ab [ABEL11].

Durch den vermehrten Einsatz von digitalen Werkzeugen und Cyber Physischen Systemen (CPS) ist ein rasanter Anstieg von industriellen Datenbeständen entlang des Fertigungsprozesses zu verzeichnen [DEUS13]. Der zunehmende Automatisierungsgrad von modernen Maschinen steigert die Verfügbarkeit von relevanten Maschinen-, Prozess- und Qualitätsdaten [HERI96]. Diese Daten können interessante und wertvolle Informationen über den Prozess enthalten, jedoch ist die Sichtung und Interpretation nur bei einfachen Zusammenhängen durch die Mitarbeiter möglich. Daten entwickeln sich zum neuen Rohstoff, der zukünftige Wettbewerbsvorteil wird von der Fähigkeit abhängen, komplexe Informationen zu konsumieren, zu produzieren und zu steuern [KING14]. Die Nutzung von modernen Informationstechnologien und Computersystemen verursacht ein rasant wachsendes Datenvolumen. Die technische Entwicklung hat die Kapazität von Speichersystemen in Datenbanken enorm erhöht. Dies hat zur Folge, dass immer höhere Anforderungen an die Verarbeitung und Auswertung von Informationen gestellt werden. Bislang sind direkte Anfragen nach Informationen an Datenbanksystemen problemlos möglich, jedoch ist das Aufspüren von interessanten Informationen wie Muster, Strukturen und Gesetzmäßigkeiten nicht möglich. Im Rahmen der Wissensgewinnung haben sich verschiedene Data-Mining Auswertungsmethoden etabliert. Data-Mining hat das Ziel, Informationen aus den angefallenen Daten zu gewinnen und neues Wissen über die Umwelt zu erhalten [SHAR13]. Die erfassten Bestands- und Transaktionsdaten aus Lager, Maschinen, Werkzeugen und Qualitätskontrollen, können durch Methoden des Data-

Mining verwertet und zur Wissensgewinnung zurückgeführt werden und somit zur Optimierung der Fertigungsprozesse beitragen. Im Produktionsumfeld werden Data-Mining Methoden recht zurückhaltend eingesetzt obwohl großes Potenzial bezüglich der Prozessoptimierung möglich ist und klassische Methoden ergänzt werden können [BERN11].

Das Ziel dieser Arbeit ist die Untersuchung der Wirkzusammenhänge der Fertigungsprozesse und Prozessparameter auf die Bauteilqualität anhand von Produktionsdaten. Hierzu werden die erforderlichen Grundlagen bezüglich Produktion, Qualitätsmanagement, KDD und Data Mining aufgearbeitet. Des Weiteren erfolgt die Darstellung und Erläuterung relevanter KDD-Vorgehensmodelle und Data Mining Methoden. Anschließend wird ein geeignetes KDD-Vorgehensmodell und passende Data-Mining Methode in Bezug auf die Fertigung von elektronischen Bauteilen ausgewählt. Es folgt eine praktische Umsetzung im Rahmen einer Fallstudie sowie Ableitung konkreter Handlungsempfehlungen.

2.Grundlagen

2.1 Produktion und Fertigung

Die Produktion hat das Ziel, Güter durch den Einsatz von Produktionsfaktoren, in der richtigen Menge und Qualität zum richtigen Zeitpunkt und zu minimalen Kosten zu fertigen [BUNG12]. In diesem Abschnitt wird ein Einblick auf die Themenfelder Fertigung, Qualitätsmanagement und Informationssysteme gegeben, um einen Einblick in die Produktion zu geben.

2.1.1 Begriffliche Grundlagen

Im Folgenden wird die Produktion, anhand von grundlegenden Begriffen, in die systemtechnische und prozessorientierte Sichtweise eingeordnet. Zunächst wird der Begriffe „System“ und „Prozess“ definiert.

System und Prozesse

Ein System ist eine vereinfachte Darstellung eines Ausschnitts der Realität und wird verwendet um Fragestellung näher zu untersuchen. Es besteht aus Systemelementen und Vernetzungen zwischen den einzelnen Elementen, welches die Funktion und Struktur bestimmt. Um die Wechselbeziehungen des Systems zu präzisieren, werden Systemgrenzen zur Umwelt definiert, die aufgrund der Komplexität nicht in allen Fällen leicht zu bestimmen sind. Werden Wechselwirkungen des Systems mit der Umwelt berücksichtigt, handelt es sich um ein offenes, andernfalls um ein geschlossenes System. Die Zusammenfassung mehrerer Systeme führt zu einem Supersystem. Durch Untergliederung entstehen Subsysteme, wodurch die Betrachtungsweise und Analyse vereinfacht wird [BOSS04].

Diese Arbeit beschäftigt sich schwerpunktmäßig mit dem „System Produktion“ (vgl. **Abbildung *1***), daher folgt eine kurze und prägnante Erläuterung der Begrifflichkeiten. Der Begriff Produktion definiert sich nach Hachtel und Holzbaur wie folgt, „Transformation von Ausgangsstoffe in Ausbringungen“ [HACH10]. Das System Produktion besteht aus mehreren Elementen, die für die Wertschöpfung relevant sind. Als Elemente kommen Menschen, Maschinen, Anlagen, technische und organisationale Prozesse zum Einsatz. Um jedes Element nochmal einzeln zu untersuchen, werden diese in Subelemente unterteilt. Durch die vorgenommene Unterteilung ist es möglich, Wechselwirkungen auf Mikro - Makro Ebene zu betrachten. Durch die Wechselwirkung innerhalb des Systems, können Veränderungen an den Elementen, wie z.B. Optimierung von Prozessen, Gestaltung von Arbeitsplätzen, Auswirkungen auf das gesamte System haben. Maßgebliche Einflussfaktoren, die die Systemgrenze der

industriellen Produktion beeinflussen, werden in innerliche und äußerliche Einflüsse eingeteilt. Die inneren Einflüsse entstehen durch die Organisation, den Ressourcen und der Veränderung der Fähigkeiten. Die äußeren Einflüsse sind hingegen nicht beeinflussbar, da sie marktgetrieben sind [WEST16].

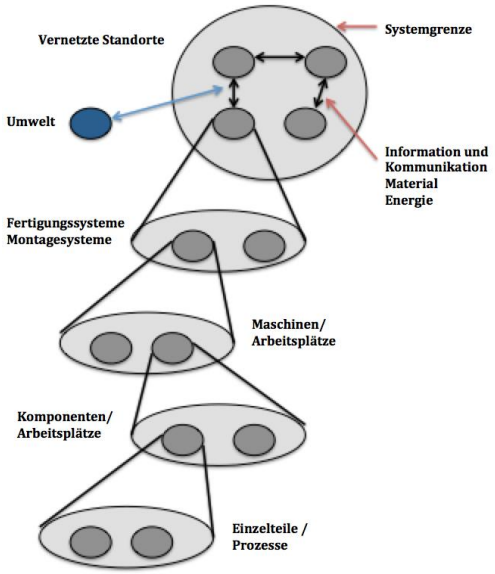


Abbildung 1: Das System Produktion nach [WEST16]

Unternehmen erzeugen Produkte, um Kundenwünsche und -bedürfnisse zu befriedigen, mit dem Ziel den wirtschaftlichen Erfolg weiterhin zu sichern. In Prozessen werden die Leistungen erstellt. Unter einem Prozess ist eine „zeitabhängige Zustandsänderung in einem System“ [SCHM02] zu verstehen, die Eingangsgrößen (Input) in Ausgangsgrößen (Output) transformiert (vgl. **Abbildung 2**). Mit den Anforderungen an die Prozesse ist oft der Produktrealisierungsprozess bzw. Herstellungsprozess gemeint. Die Gestaltung der jeweiligen Prozesse nimmt wesentlichen Anteil an der Qualität des Endproduktes [HERR16].

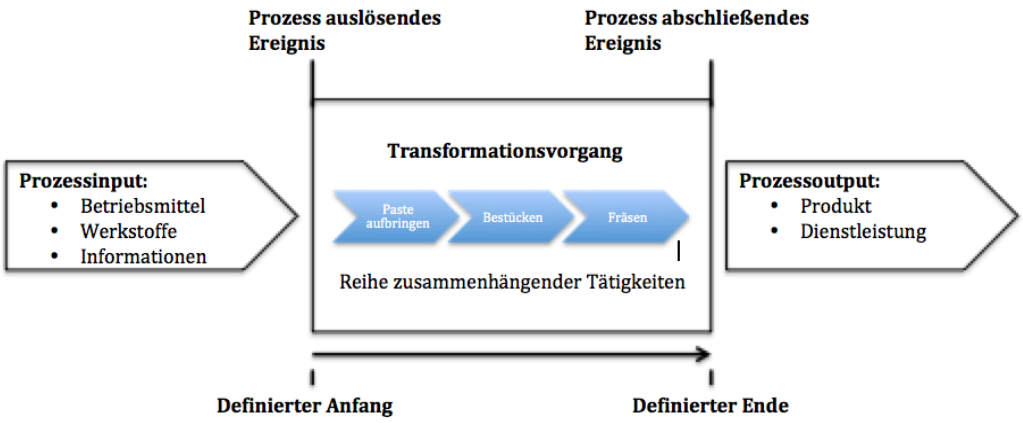


Abbildung 2: Schematische Darstellung eines Prozesses nach [SCHM02]; [GÖTZ13]

Die Eingangsgrößen (Input) bilden Einsatzfaktoren wie Betriebsmittel, Werkstoffe und Informationen, die auf den Prozess einwirken, ohne selbst von ihm beeinflusst zu werden. Die Ausgangsgröße (Output) werden als Größen eines Prozesses genannt, welche von den Eingangsgrößen sowie vom Prozess selbst beeinflusst werden und liegen in materieller Form - als Produkt - oder immaterieller Form - als Dienstleistung - vor. Die Eingangsgrößen werden rechtzeitig zum Prozessbeginn vom Lieferanten bereitgestellt. Der Output ist für den Kunden bzw. für die nachfolgenden Prozesse bestimmt. In diesem Zusammenhang sprich man von Kunden-Lieferanten-Beziehung, beide stellen Qualitätsanforderungen an Eingangs- und Ausgangsgrößen [SCHM02].

2.1.2 Geschäftsprozesse in Industrieunternehmen

Nachdem nun System und Prozesse definiert worden sind, wird nun das Augenmerk auf den Begriff Geschäftsprozess gelegt.

Die Begriffe „Prozess“ und „Geschäftsprozess“ werden in der Praxis sowie in der Literatur häufig synonym verwendet. Allerdings stellt der Geschäftsprozess eine bestimmte Prozessart dar, der Begriff Prozess hingegen umfasst alle Arten von Prozessen. Nach Schmelzer und Sesselmann grenzt sich der Geschäftsprozess vom allgemeinen Prozess - durch die Fokussierung auf den Kunden - ab. Die Autoren Schmelzer und Sesselmann definieren den Geschäftsprozess als eine „funktions- und organisationsübergreifende Folge aus wertschöpfende Aktivitäten“, welche die Anforderung vom Kunden in eine für den Kunden geschaffene Leistung umwandelt. Der Geschäftsprozess beginnt und endet beim Kunden [SCHM13].

Geschäftsprozesse können laut Schmelzer und Sesselmann in primäre und sekundäre Geschäftsprozesse unterschieden werden. Der primäre Geschäftsprozess wird direkt vom externen Kunden angestoßen. Innerhalb des Prozesses werden Leistungen für den Kunden erzeugt und bilden somit den Kern der Wertschöpfung. Die sekundären Geschäftsprozesse beinhalten Unterstützungsaktivitäten für den effektiven und effizienten Ablauf von primären Prozessen. Als unterstützende Aktivität ist die Bereitstellung von bedarfsgerechten Ressourcen sowie der nötigen Infrastruktur zu verstehen. Der sekundäre Geschäftsprozess liefert somit einen indirekten Beitrag zur Leistungserbringung, die vom externen Kunden angestoßen werden. Da es sich bei sekundären Geschäftsprozessen um interne Kunden handelt, besteht somit kein direkter Marktbezug. Für die Bewertung interner Leistungen sollten die gleichen Maßstäbe, wie bei externe Leistungen gelten [SCHM13].

Der Schwerpunkt dieser Arbeit liegt auf der industriellen Fertigung zur Herstellung von Stückgütern. Daher sind folgende primäre Geschäftsprozesse von besonderer Bedeutung: Produktentstehungs- und Auftragsabwicklungsprozess (vgl. **Abbildung 3**) [GRÖG15].

- Der Produktentstehungsprozess umfasst die Teilprozesse der Produktentwicklung und Produktionssystementwicklung. Die Grundlagen für die Produktideen werden aus den Anforderungen des Kunden abgeleitet.
- Der Auftragsabwicklungsprozess umfasst die Planung und Steuerung sämtlicher Material- und Informationsflüsse vom Kundenauftragseingang über die Herstellung bis zur Distribution des Produktes an den Kunden. Dieser Prozess besteht aus mehreren Teilprozessen.

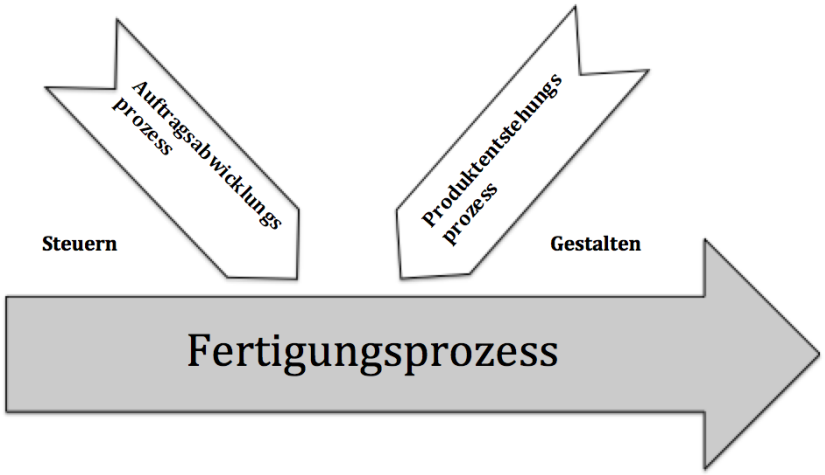


Abbildung 3: Zusammenhang von Geschäfts- und Fertigungsprozessen nach [GRÖG15]

2.2 Qualität und Qualitätsmanagement

Der Begriff **Qualität** wird nach der DIN EN ISO 8402 (1995) und der American Society for Quality Control (ASQC) folgendermaßen definiert [DIN95]:

„Qualität ist die Gesamtheit der Merkmale und Merkmalswerte eines Produktes oder einer Dienstleistung bezüglich ihrer Eignung, festgelegte und vorausgesetzte Erfordernisse zu erfüllen“

Von dieser Definition ausgehend müssen alle Eigenschaften und Merkmale eines Produktes den Anforderungen entsprechen. Die Nichterfüllung der Anforderung beim Soll-Ist-Vergleich wird als Fehler gewertet und somit als fehlerhaftes Produkt deklariert. In der Norm DIN EN ISO 9000 (2015), die international als Grundlage Qualitätsmanagementsysteme dient, wird Qualität abstrakter definiert [DIN15].

„Grad, in dem ein Satz inhärenter Merkmale Anforderungen erfüllt“

Inhärente Merkmale bezeichnen folglich kennzeichnende Eigenschaften eines Produktes, Systems oder Prozesses zur Erfüllung von Kundenanforderungen und anderer interessierter Parteien. Hierbei ist das Augenmerk nicht nur auf den Endkunden ausgerichtet, es werden auch

Lieferanten, Politik und Gesellschaft mit einbezogen. Das Produkt muss nicht nur Kundenerwartungen erfüllen, sondern auch behördliche Anforderungen, wie z.B. in Sicherheit- und Umweltrichtlinien gefordert [KIEH01]; [SCHM15]. Die Komplexität und Vielschichtigkeit des Qualitätsbegriffs werden aufgrund der unterschiedlichen Definitionsansätze deutlich und unterliegen einer ständigen Veränderung [GÖTZ13]. Die Wandlung der Sichtweise wird nochmal anhand von **Abbildung 4** verdeutlicht.

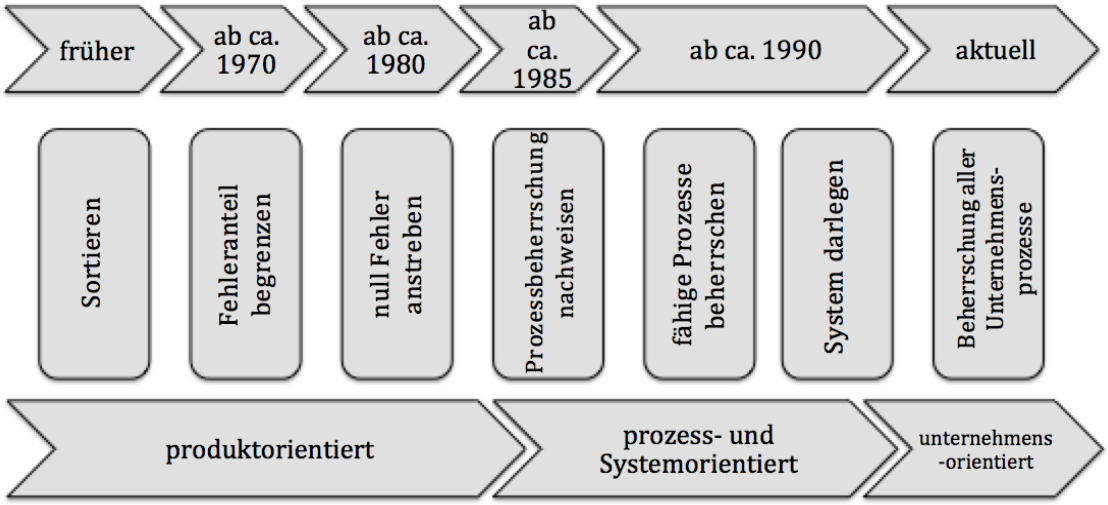


Abbildung 4: Wandel des Qualitätsverständnisses nach [Pfei14]

Mit dem Beginn der Produktion ist die Qualität der hergestellten Produkte stets ein wichtiges und viel diskutiertes Thema. Vor 1970 war das Qualitätswesen ein reines Kontrollinstrument, welches zur Erfüllung der Anforderung an das jeweilige Produkt eingeführt worden ist. Man spricht von einer produktorientierten Kontrolle, welche sich zur damaligen Zeit auf das Sortieren, Inspizieren und Reparieren beschränkt hat. Der Anspruch ändert sich zunehmend, sodass im Laufe der Zeit nicht mehr nur von der Endkontrolle, sondern von Prozesskontrolle gesprochen wird. Die Ursachen für fehlerhafte Produktion ist untersucht und verbessert worden, um ein wiederholtes Auftreten der Fehler zu verhindern. Dieses hatte zur Folge, dass das Qualitätsbewusstsein nicht mehr nur auf der operativen Ebene beschränkt war, sondern auch im Führungsmanagement vertreten war [SCHM15].

Das moderne Qualitätsmanagement (QM) hatte sein Ursprung in Japan entwickelt. Es soll zielführend in Produktionsbetrieben eingesetzt werden. Die Optimierung von Produktions- und Dienstleistungsprozessen gestaltet sich zum entscheidenden Faktor, um auf dem globalisierten Markt bestehen zu können. Die Tätigkeiten umfassen die Festlegung von Qualitätszielen, -verbesserung, -maßnahmen und -verantwortung. Im QM werden drei verschiede Objekte, auch Betrachtungsebenen genannt, fokussiert: Produkte und Dienstleistung sowie Prozesse und Systeme (vgl. **Abbildung 5**).

Die Anforderungen werden vom potenziellen Kunden an das Produkt - mit entsprechenden technischen Zeichnungen, Tabellen und Mustern - gestellt. Für die Herstellung des Produktes, muss das QM für jeden einzelnen Bestandteil die inhärenten Merkmale genau festlegen. Des Weiteren stellt das QM sicher, dass die Gestaltung der Arbeitsabläufe im Betrieb die qualitativen Anforderungen an die Produkte erfüllt. Die Gestaltung des Herstellungsprozesses beeinflusst maßgeblich die Qualität des hergestellten Produktes [SCHN08]. Um die Produkt- und Prozessqualität zu gewährleisten, muss das Qualitätsmanagementsystem die vier Hauptprozesse wie Planen, Durchführen, Prüfen und Handeln erfüllen sowie die benötigten Ressourcen wie Personal, Infrastruktur, Messsysteme und Wissen bereitstellen [HERR16]; [SCHN08].

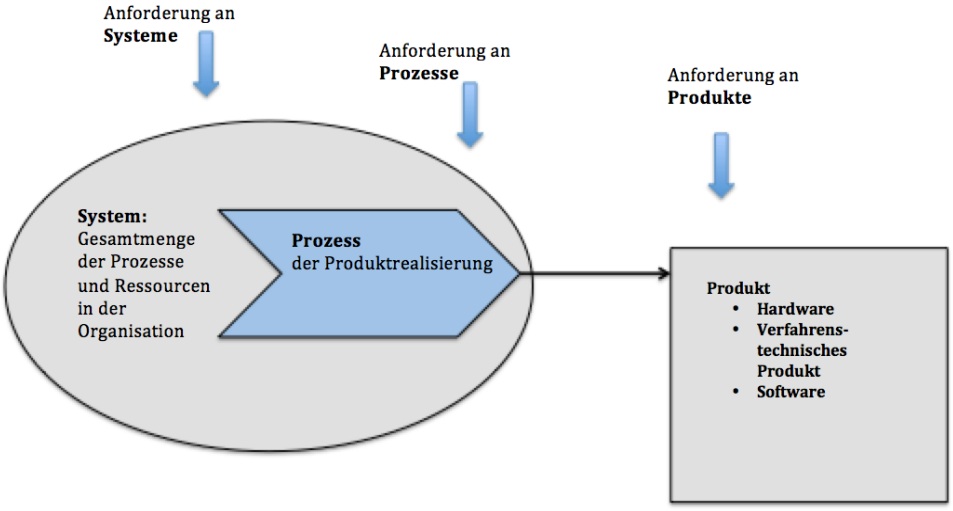


Abbildung 5: Betrachtungsebenen QM nach [HERR16]

2.2.1 Ziele im Qualitätsmanagement

Die hohe Produktqualität ist die Grundlage für den wirtschaftlichen Erfolg eines Produktionsbetriebs. Die Produktqualität muss mindestens den Kundenanforderungen entsprechen oder diese bestenfalls übertreffen, um eine hohe Kundenzufriedenheit und -bindung zu erzielen. Die dauerhafte Produktion bzw. Lieferung von minderer Qualität wird auf Unzufriedenheit stoßen und birgt die Gefahr, den Kunden dauerhaft zu verlieren. Des Weiteren ist die reine Erfüllung des Kundenwunsches nicht ausreichend, dies wird als Grundvoraussetzung gewertet und erzeugt keine besondere Begeisterung beim Kunden. Anhand des Kano-Modells werden die Zusammenhänge dargestellt (s. **Abbildung 6**). Die Basisfaktoren werden vom Kunden als selbstverständlich empfunden und werden nicht direkt artikuliert. Das Nichtvorhandensein der Basisfaktoren ruft beim Kunden Unzufriedenheit hervor. Einen maßgeblichen Einfluss haben Leistungsmerkmale, auch Qualitätsmerkmale genannt, diese werden vom Kunden explizit verlangt. Für eine erhöhte Zufriedenheit sorgen

Begeisterungsmerkmale, die jedoch vom Kunden nicht erwartet werden. Die Merkmale sind zeitabhängig, die jetzigen Leistungsanforderungen werden im Laufe der Zeit zu Basisanforderungen. Die Aufgabe des Qualitätsmanagements ist es, die Basis - und Leistungsanforderungen des Kunden im Bereich Qualität zu erfüllen [GAUB09].

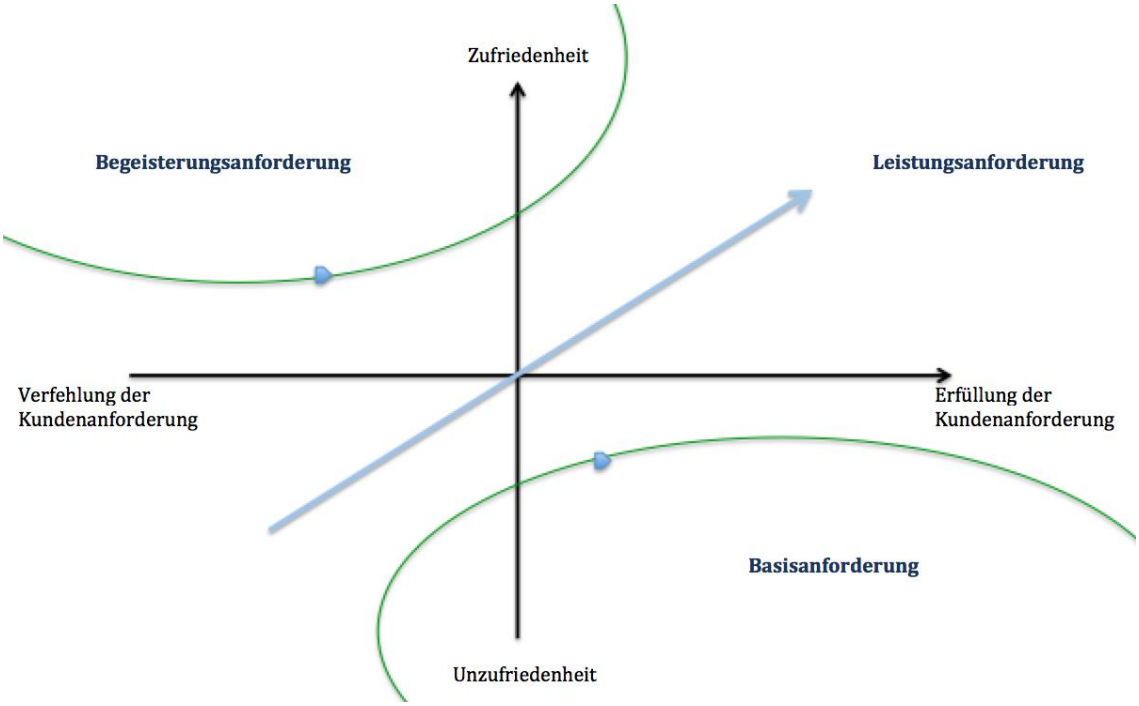


Abbildung 6: Kano-Modell nach Kano [GAUB09]

2.2.2 Qualitätssicherung

Qualitätsfähigkeit

Bei der Entwicklung und Auslegung von Produktionsprozessen ist die Gewährleistung der Produktqualität das oberste Ziel. Für die Bewertung der Qualität eines Produktes werden Qualitätsmerkmale definiert, deren Ausprägung entscheidend ist. Um die Anforderung an das Produkt zu erfüllen, wird für jedes Qualitätsmerkmal die Ausprägung festgelegt. Die Ausprägung, auch Sollausprägung genannt, muss vorliegen, um das Produkt in den geforderten Angaben realisieren zu können. Der Verbraucher kann damit der Endkunde oder der Mitarbeiter des nachfolgenden Prozesses sein, der die Bewertung des Produktes durchführt und ermittelt, ob die Ausprägung der Qualitätsmerkmale den vorgegebenen Kriterien entsprechen. Bei bestimmten Qualitätsmerkmalen muss die Sollausprägung exakt eingehalten werden, bei anderen wiederum orientiert man sich an dem Toleranzbereich. Der Toleranzbereich ist mit Mindest- und Höchstausprägung definiert und erlaubt Abweichungen von der Sollausprägung, bis zu einem fest

definierten Maß. Liegt die Istausprägung innerhalb des Toleranzbereichs, gelten die Merkmale als erfüllt und werden als fehlerfrei bewertet. Außerhalb des Toleranzbereiches ist das Produkt nicht qualitätskonform und verursachen Kosten, z.B. durch Nachbearbeitung oder Ausschuss [BRÜG12].

Von der traditionellen Vorstellung abgewandert, zeigt uns die von Taguchi entwickelte Qualitätsverlustfunktion (vgl. **Abbildung 7**), dass es nicht unerheblich ist, wie weit die Istausprägung von der Sollausprägung im Toleranzbereich abweicht. Jegliche Abweichung der Sollausprägung wird als ansteigender Qualitätsverlust bewertet und verursacht Kosten. Die Verlustfunktion verfolgt das Ziel der vollständigen Erfüllung der Kundenanforderung und steht im Einklang mit der „Null-Fehler-Strategie“ [BRÜG12].

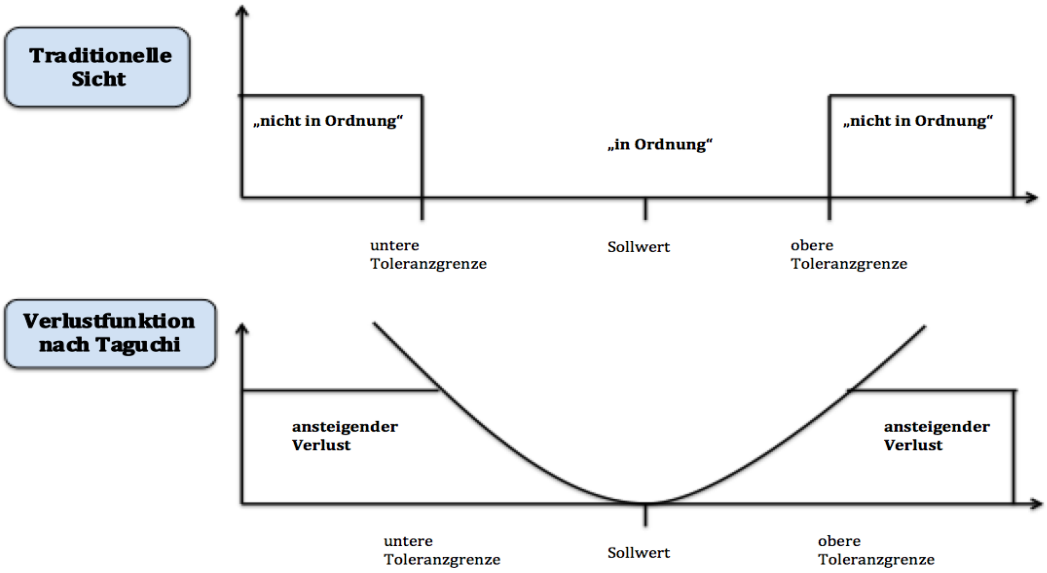


Abbildung 7:Verlustfunktion nach Taguchi [BRÜG12]

Prozessfähigkeit

Um Produktionsprozesse bezüglich ihrer Fähigkeit, Produkte mit qualitätskonformen ausgeprägten Merkmalen hervorzubringen, wird der Begriff Prozessfähigkeit geprägt. Die Prozessfähigkeit wird anhand qualitativer und quantitativer Prozessfähigkeitsanalyse bestimmt. Die qualitative Analyse erfolgt durch die Beobachtung von Merkmalen (z.B. glänzende oder matte Oberfläche) und der Prozess wird anhand einer ppm-Rate bewertet [JUNG13].

$$ppm = \frac{Anzahl\ der\ defekten\ Einheiten}{Anzahl\ der\ Einheiten} * 1000000$$

Die quantitative Analyse, auch Statistische Prozesskontrolle (SPC) genannt, erfolgt durch Messung von Merkmalen einer statistisch repräsentativen definierten Anzahl von Prozessergebnissen. Für die im Rahmen dieser Produktionsprozesse erzeugten Qualitätsmerkmale, werden nun einzeln für jedes Merkmal Verteilungsfunktionen ermittelt, um die Streuung der Qualitätsausprägung um den Sollwert darzustellen. Anhand der Ergebnisse werden die Streubreite und die Lage der Verteilungsfunktion ermittelt. Für die Bewertung der Fähigkeit des Produktionsprozesses werden schließlich Streubreite und Lage der Verteilungsfunktion mit der Toleranzgrenze verglichen. Die Toleranzgrenze wird durch den Kunden vorgegeben, dafür werden Toleranzgrenzen (UGW und OGW) festgelegt (vgl. **Abbildung 8**). Ein Produktionsprozess, der Produkte mit qualitätskonform ausgeprägten Qualitätsmerkmalen herstellen, wird als qualitätsfähig bezeichnet. Die Prozessfähigkeitsuntersuchung legt die Grundlage für die Qualitätsfähigkeit der einzelnen Produktionsprozesse. Anhand des Prozessfähigkeitsnachweises wird die Qualitätsfähigkeit der Produktionsprozesse gewährleistet [JUNG13].

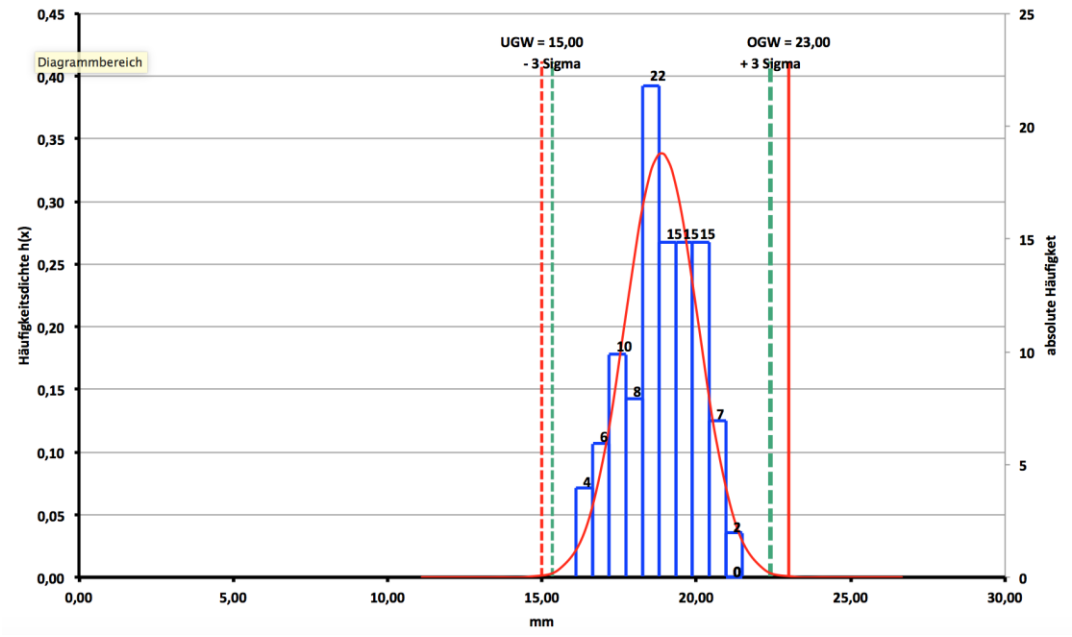


Abbildung 8: SPC

2.2.3 Qualitätsregelung

Die Umsetzung qualitätsrelevanter Maßnahmen in der Fertigung liegt im Aufbau von Regelkreisen, betriebliche Zusammenhänge werden in Anlehnung an technische Regelkreise modelliert. Die Regelungstechnik beschäftigt sich mit der Modellierung der realen Verhaltensweisen des zu regelnden Systems in einen technischen Regelkreis. Die

Zusammenhänge in der Produktion können durch die Systematik des technischen Regelkreises vereinfacht und visualisiert dargestellt werden. Um einen gewünschten Zustand in einer Regelstrecke zu erreichen und zu halten, werden Regelelemente eingesetzt. In Regelkreissystemen werden Regelgrößen durch kontinuierliche Messungen auf den vorgegebenen Sollwert gehalten. Dementsprechend werden im Qualitätsregelkreis, die zu regelnde Regelgröße -Qualität- mit dem Einsatz von Regelelementen auf dem vorgegebenen Sollwert gehalten (s. **Abbildung 9**) [OETZ05].

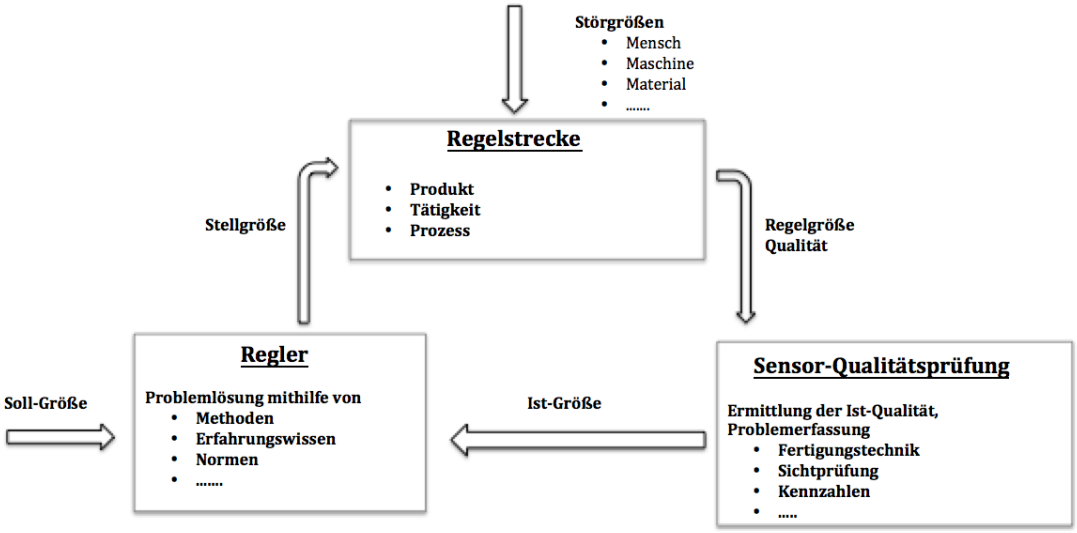


Abbildung 9: Qualitätsregelkreis nach [OETZ05]

Die Sollgröße stellt die Qualitätsanforderung im Qualitätsregelkreis dar. Die Regelgröße, in dem Fall auch Sollgröße, geht als Eingangsgröße in die Regelung ein. Im Regler wird die Abweichung zwischen Soll- und Istgröße berechnet und damit die Stellgröße abgeleitet, um die vorgegebenen Qualitätsanforderungen zu erfüllen. Im Regler kommen Qualitätsmethoden wie z.B. SPC zum Einsatz um die Abweichung zu beheben. Die Regelstrecke eines Qualitätsregelkreises ist im klassischen Sinne ein Prozess z.B. eine Maschine, dessen Ausgangsgröße geregelt wird. Hierbei wird der Prozessbegriff im Vergleich zum technischen Regelkreis, auf alle Tätigkeiten, die Einfluss auf die Produktentstehung nehmen z.B. Konstruktion, Planung, Fertigung oder Montage, erweitert. Bei der Umsetzung der Qualitätsmaßnahmen durch die Stellgröße wirken auf die Regelstrecke Störgrößen, die den Sollzustand erschweren, ein. Als Störgrößen kommen dabei Größen wie z.B. Mensch, Maschine, Material, Umwelt in Frage [OETZ05].

2.2.4 Qualitätsmangel

Von einem Qualitätsmangel spricht man, wenn bei der Herstellung eines Produktes, die im Sollzustand festgelegten Merkmale nicht ausreichend erfüllt worden sind. Ein Qualitätsmangel ist in erster Linie auf die nicht qualitätsfähigen Produktionsprozesse zurückzuführen. Wenn ein

Prozess Qualitätsanforderungen nicht erfüllen kann, erzeugt er mangelhafte Produkte, somit entstehen dem Unternehmen zusätzliche Kosten und Reklamationen, bis hin zu Konventionalstrafen. Das Produkt kann aufgrund seines Qualitätsverlustes nicht an die nachfolgenden Prozesse weitergeben bzw. an den Kunden verkauft werden. Im Rahmen der betrieblichen Nachbearbeitung wird versucht, die Ausprägung des Qualitätsmerkmals den Kundenanforderungen anzupassen, sollte dies nicht möglich sein, ist das Produkt als Ausschuss zu bewerten und muss entsorgt werden. Falls mangelhafte Produkte in den Wirtschaftskreislauf gelangen, kann das Image des Unternehmens enorm geschädigt und Kunden dabei verletzt werden [HEPP08].

Bei einem Produktionsprozess, der kontinuierlich mangelhafte Produkte hervorbringt, ist die Prozessfähigkeit nicht gegeben. Um die Ursache der mangelnden Qualitätsfähigkeit zu ermitteln und zu beheben, muss ein hoher Aufwand der Fehleranalyse, welche erneut Kosten aufwirft, betrieben werden. Kann während der Ursachenanalyse nicht produziert werden, entstehen dem Unternehmen zusätzliche Belastungen. Aufgrund des außerplanmäßigen Stillstands kommt es zu einer Reorganisation der Produktionsplanung und somit zu Lieferengpässen. Soll der Liefertermin dennoch eingehalten werden, müssen die Ausfallmengen durch zusätzliche Ressourcen in den Bereichen Personal, Material und Maschine aufgefangen werden, um die geforderte Kundenmenge rechtzeitig zu liefern [HEPP08]; [JUNG13].

2.3 Daten und Informationsmanagement in der Produktion

2.3.1 Informations- und Kommunikationstechnik

Die Grundlage für ein effizientes Informationsmanagement - bezogen auf die Produktion - bilden moderne Informations- und Kommunikationstechnologien (IuK). Sie beschleunigen die Arbeitsteilung zwischen internen Standorten und externen Unternehmen [WEST06]. Ein typischer Vertreter produktionsnaher IT-Systeme ist das Manufacturing Execution System (MES). Es ist in der Fertigungsebene angesiedelt und unterstützen das Fertigungsmanagement. Eine der Aufgaben ist das Abdecken der Informationslücken zwischen der Unternehmensleit- und Fertigungsebene, um z.B. kurzfristige Änderung in der Feinplanung und Produktionssteuerung zu ermöglichen. Gemäß VDI-Richtlinien 5600 (2012) können maximal zehn Aufgabenbereiche der MES zugeordnet werden (vgl. **Abbildung 10**). Bezogen auf ihre Funktionalität kann die MES in drei Gruppen - dem Qualitätsmanagement, Fertigungsmanagement und Personal - zusammengefasst werden. Je nach Unternehmensausrichtung können die Aufgabenbereiche ausgesucht und, dem für die jeweilige Produktion benötigten Leistungsmodule, angepasst werden. Die Aufgabenmodule sind unterhalb der Unternehmensleitebene und oberhalb der Fertigungsebene angesiedelt. Eine klare Teilung der einzelnen Ebenen ist nicht möglich, da die

Bereiche ineinander übergehen und Informationen austauschen. Die vertikale Integration der Ebenen und Systeme ermöglicht den Informationsaustausch, ein MES empfängt und verarbeitet Daten aus der Fertigung und tauscht gleichzeitig Informationen mit dem ERP-System aus [KLET15]; [VDI13].

Einer der wichtigsten Aufgabe des MES ist die Überwachung und Gestaltung der Produktion in Echtzeit. Die relevanten Daten einer Produktion werden, wenn möglich, vollautomatisch durch Messsysteme, Maschinen, Barcodeleser, RFID und über weitere Schnittstellen erfasst. Die Übertragung der Daten in Echtzeit ermöglicht den Verantwortlichen ein schnelles Reagieren auf Veränderungen oder Schwachstellen in den Prozessabläufen. Dadurch kann ein frühzeitiges Eingreifen in den Fertigungsprozess gewährleistet werden, um zusätzliche Produktionskosten oder diese im Zuge der Prozessoptimierung zu verringern. MES agiert technologieorientiert, zeitnah und gibt die tatsächlichen Verhältnisse im Produktionsbereich wieder [KURB16].

Alle drei Ebenen und Systeme arbeiten mit unterschiedlichen Zeithorizonten. Um mit den zugeschickten Informationen arbeiten zu können, ist ein zeitgerechter Datenaustausch essentiell. Des Weiteren ist die Bereitstellung „sinnvoller“ Daten für die überlagerte und unterlagerte Ebene vorzunehmen. Die Auswahl der Daten kann individuell an die Unternehmensziele angepasst werden. Ein ERP-System braucht beispielsweise keine Informationen über mehrmaligen Produktionsstillstand oder Maschinenzustand. Das MES als - zwischen geschaltetes System - ermöglicht die exakte Erfassung und Kommunikation unter den Ebenen[KLET07].

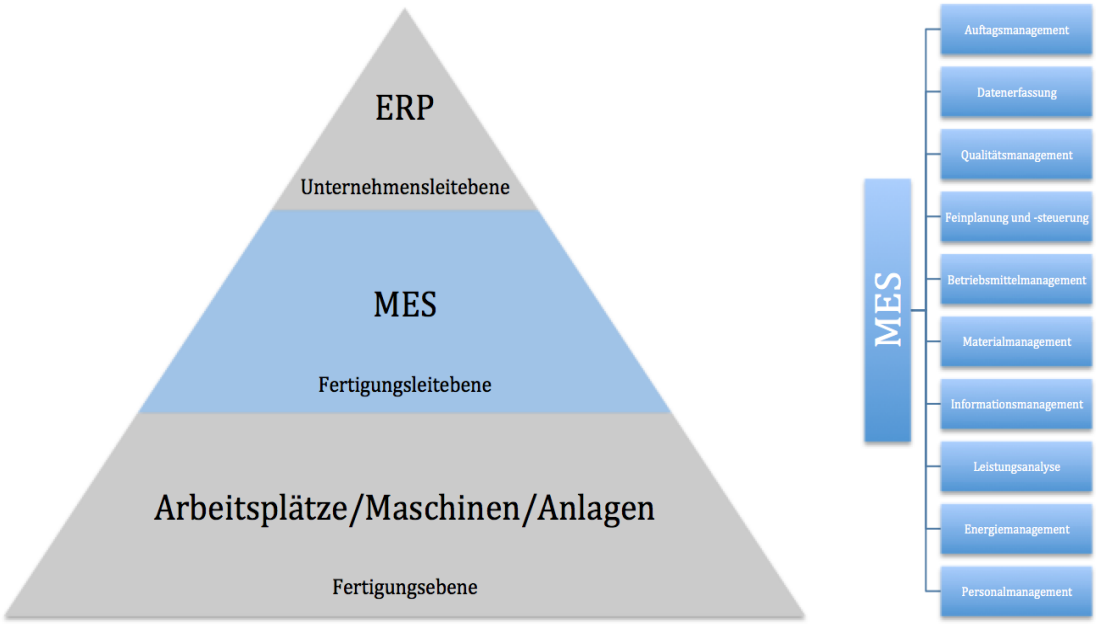


Abbildung 10: Anwendungssystemarchitektur nach *[KLET15]*

2.3.2 Betriebsdatenerfassung

Die Betriebsdatenerfassung (BDE) gehört zu den wichtigsten Kernfunktionalitäten des MES. Die BDE umfasst alle Maßnahmen um Betriebs- und Maschinendaten zu erfassen und schließlich in maschinell lesbarer Form wiederzugeben. Betriebsdaten sind Informationen, die täglich in der Fertigung und den Prozessketten anfallen, sie bilden somit eine wichtige Grundlage für Informations- und Auswertungssysteme. Die Fertigungsmaschinen sind für Industrieunternehmen der wichtigste Faktor zur Leistungserbringung. Um sie wirtschaftlich zu nutzen, ist eine effektive Auslastung der Maschine essentiell. Dies kann nur erfüllt werden, wenn umfassende Informationen über die Fertigungsanlagen rechtzeitig vorliegen, um Störungen und Fehlentwicklungen zu vermeiden. Anhand der BDE können Maschinedaten lückenlos erfasst, visualisiert und ausgewertet werden und somit zur Produktivitätssteigerung und gleichzeitigen Kostensenkung beitragen.

Hierbei ist zu beachten, dass BDE als Oberbegriff für Erfassungsverfahren wie Maschinendatenerfassung, Auftrags-, Prozess-, Qualitätsdatenerfassung, Personalzeiterfassung usw. dient. Der genaue Leistungsumfang der BDE hängt von den jeweiligen Anforderungen des Unternehmens ab und kann entsprechend zusammengestellt werden [KOUK01].

Erfassung der Betriebsdaten

Die Erfassung der Betriebsdaten erfolgt am Entstehungsort bzw. auf der Erfassungsebene. Informationen werden IT-gestützt in Echtzeit in der Produktion erfasst. Hierzu bieten sich verschieden Möglichkeiten an [KLET07].

- Automatische Erfassung von Prozessdaten
  Die Prozessdaten können an den Maschinen, durch den Einsatz von Sensoren, automatisch erfasst werden. Die Informationserfassung an der Anlage wird auch als Maschinendatenerfassung (MDE) bezeichnet. Prozessdaten umfassen sowohl die notwendigen Informationen für die Inbetriebnahme (Soll-Daten), als auch die direkt im Betrieb erzeugten Informationen (IST-Daten). Dabei werden die wesentlichen Maschinendaten wie z.B. Temperatur, Geschwindigkeit, Zeit, Gewicht usw. aufgezeichnet.

- Identifizierungssysteme
  Der Einsatz von Datenträger auf Objekten ermöglicht die Erfassung der Informationen über Lesegeräte und vermeidet somit die manuelle Eingabe. Die Datenerfassung erfolgt anhand von Barcode oder RFID.

- Mobile Erfassung

  Mit Hilfe von mobiler Datenerfassung werden Informationen ortsungebunden erfasst. Die Erfassungsgeräte können je nach Bedarf mit einem Barcodescanner oder RFID - Technik ausgestattet sein.

Die Betriebsdatenerfassung unterscheidet verschiedene Arten von Betriebsdaten [KLET15]:

- Organisatorische Betriebsdaten
  - Auftragsdaten
    - Produktionsdaten wie Zeit, Anzahl, Stückzahl
    - Arbeitsfortschritt, Auftragsstatus
  - Personaldaten
    - Anwesenheit- und Arbeitszeit
    - Zutrittskontrolle
    - Lohnkosten

- Technische Betriebsdaten
  - Maschinendaten
    - Laufzeiten und Unterbrechungen von Maschinen
    - Meldungen von Störungen
    - Daten der Instandhaltung
    - Messwerte, Temperatur, Druck, Geschwindigkeit
    - Verbrauch von Material, Energie und Hilfsmittel
  - Prozessdaten
    - Qualität
    - Parameter der Prozesse
    - Einstelldaten

3. KDD als Werkzeug der Prozessanalyse

Im Rahmen dieses Kapitels werden die erforderlichen Grundlagen von Knowledge Discovery in Database und des Data-Mining anhand grundlegender Begriffe näher erläutert und die Bedeutung des KKD-Einsatzes in der Produktion verdeutlicht. Darüber hinaus werden ausgewählte Vorgehensmodelle für die Durchführung der Wissensgewinnung sowie Data-Mining-Verfahren vorgestellt. Der Prozess Datenaufbereitung wird aufgrund seiner Bedeutung ausführlich erläutert.

3.1 Knowledge Discovery zur Wissensgewinnung

Der industrielle Digitalisierungswandel aller Unternehmensprozesse führt zu einem stetigen Anstieg der Datenbestände. Die manuelle Untersuchung von Daten ist zeit- und kostenintensiv, die Fähigkeit die Daten zielführend zu verarbeiten und auszuwerten nimmt mit dem zunehmenden Informationsfluss ab [DEUS13]; [SHAR13]. Um bei großen Datenbeständen durch gezielte Datenanalyse Wissen zu erhalten, ist der Ansatz des Knowledge Discovery in Database (KDD)-Prozess entwickelt worden [FAYY96]. Im Rahmen dieses Kapitels werden die erforderlichen Grundlagen von Knowledge Discovery in Database und des KDD-Prozesses anhand grundlegender Begriffe näher erläutert und mit Beispielmodellen dargestellt. Des Weiteren wird das Data-Mining-Verfahren näher erläutert und die Analysemethoden vorgestellt.

3.1.1 Definition und Einordnung des KDD

Das Knowledge Discovery in Database ist eine junge Forschungsdisziplin, die im Jahr 1989 ihren Ursprung hat [SHAR13]. Der anwachsende Datenbestand und die sich daraus ergebende Möglichkeit Wissen zu generieren, hat dazu beigetragen, dass der Prozess mehr an Bedeutung gewinnt. Der KDD-Prozess wird dabei als „nichttriviale Prozess der Identifikation gültiger, neuer, potentiell nützlicher und schlussendlich verständlicher Muster in (großen) Datenbeständen“ definiert [FAYY96]; [SHAR13].

Die Einsatzgebiete von KDD sind in den Unternehmen breit gefächert, angefangen in der Finanzplanung sowie Beschaffung, Marketing und Vertrieb [SCHÖN16]. Als gängiges Alltagstool wird KDD im Vertrieb verwendet, um umsatzstarke Kunden zu identifizieren sowie Absatz - und Marktprognosen anhand von Daten zu erstellen. Um einen angemessenen Kreditrahmen für Kunden freizugeben, wird im Finanzsektor sowie im Controlling anhand von KDD die Kreditwürdigkeit abgefragt.

Unternehmensressource Wissen

Die zunehmende Datenflut in den Unternehmen bringen dem Begriff „Daten“ neue Bedeutungen zu. Daten allein haben keine eindeutige Definition, jedoch erfolgt eine gemeinsame Betrachtung und Differenzierung zwischen den Begriffen Zeichen, Daten, Informationen und Wissen (vgl. **Abbildung 11**). Die Begriffe sind eng miteinander verbunden und bauen hierarchisch aufeinander [APEL15]. Daten bestehen aus einer Reihenfolge von Zeichen und werden durch Syntaxregeln zu einer Aussage angeordnet. Die systematische Zusammensetzung von Daten in einem Kontext oder Problemzusammenhang bilden Informationen. Durch die vernetzte, strukturierte und kontextabhängige Zusammensetzung von Informationen entsteht Wissen [BODE06]; [PROB12]; [BODR03].

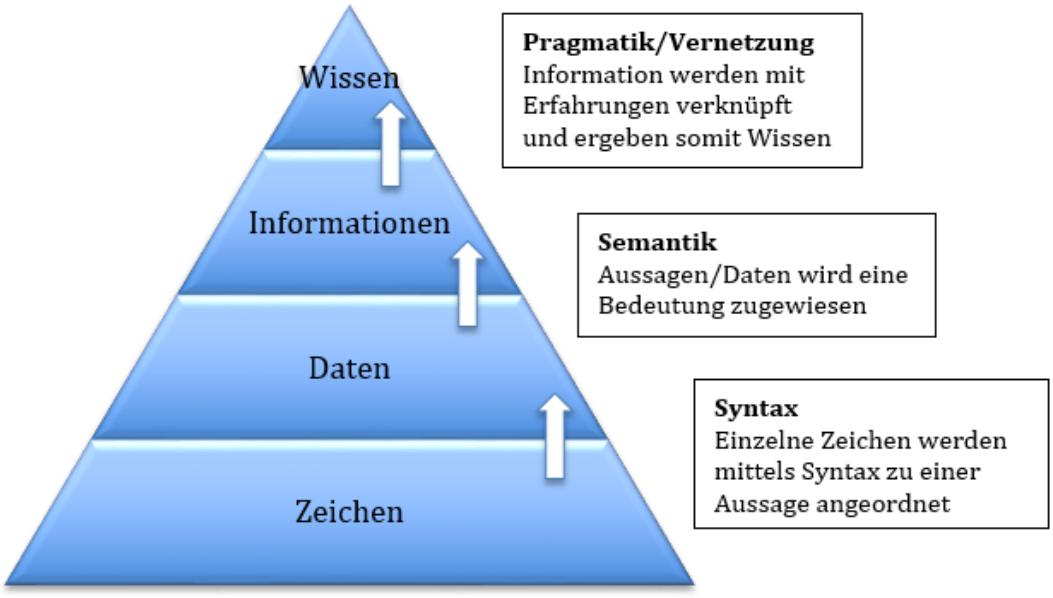


Abbildung 11: Begriffshierarchie nach [BODE06]

Wissensarten

Nach der Definition von Wissen, werden nun die Erscheinungsformen von Wissen vorgestellt [WERN04]; [BODR03].

- **Individuelles Wissen**
  Das individuelle Wissen bezieht sich auf das Wissen von Individuen und ist nur ihm selbst zugänglich. Dies ist an den einzelnen Wissensträger gebunden, welchem von Organisationen (z.B. Unternehmen) mehr Aufmerksamkeit geschenkt wird. Nur das Wissen von einem Individuum reicht nicht vollkommen aus, um Wettbewerbsvorteile zu generieren.

- **Kollektives Wissen**
  Diese Wissensart vereint das Wissen einzelner Wissensträger und verschafft sich somit eine Wissensbasis, wo wiederum jedes Individuum einzeln profitiert und sein Wissen zusätzlich erweitert. Für die erfolgreiche Umsetzung von Projekten und Strategien ist die Kombination von unterschiedlichen Wissensträgern nötig.

- **Implizites Wissen**
  Diese Form von Wissen ist dadurch gekennzeichnet, dass es schwer formulier-, kommunizier- und teilbar ist. Es ist in den Tätigkeiten und Erfahrungen des Individuums verankert und ist dem Träger, über das beherbergte Wissen, nicht bewusst.

- **Explizites Wissen**
  Darunter versteht man jenes Wissen, welches man ohne Einschränkungen abrufen kann. Es ist strukturiert, greifbar und methodisch beschreibbar und somit kann dies jedem frei zugänglich gemacht werden.

3.1.2 Die Bedeutung von Wissen im Produktionsumfeld

Die zunehmende Wissensorientierung in den Unternehmen macht Wissen zu einer wichtigen Unternehmensressource nach den bekannten Produktionsfaktoren wie Arbeit, Boden und Kapital. Ein systematischer und zielgerichteter Einsatz dieser Ressource stellt einen entscheidenden Wettbewerbs- und Entscheidungsfaktor dar. Die Methodik wird bereits erfolgreich in der Praxis bei Banken und im Vertrieb eingesetzt. Die Daten aus der Vergangenheit werden für die Analyse genutzt, um Voraussagen über die Zukunft zu treffen. Die Suche nach einem gezielten Muster, kann schließlich auf das zukünftige Verhalten der Kunden übertragen und somit können Vorsichtsmaßnahmen getroffen werden.

Bislang lassen sich folgende Vorteile durch den Einsatz der Ressource Wissen erzielen [KLOS01]:

- Wissen über Kundenverhalten ermöglicht Kundenbindung bzw. - gewinnung
- Wissen über Wettbewerber ermöglicht von ihnen zu lernen und das eigene Unternehmen zu positionieren (Benchmarking)
- Integration von Wissen schafft neue Geschäftsfelder, Prozesse und Produkte

Die Datenauswertung ist in vielen Bereichen eines Unternehmens fest im Alltag integriert, um damit weitere Potenziale auszuschöpfen. Laut Frauenhofer IOSB ist die Wissensentdeckung im Produktionsumfeld bisher nicht etabliert worden. Der Wettbewerbsdruck erfordert kostensenkende Prozesse, um weiterhin erfolgreich am Markt bestehen zu können. Die Komplexität von Produktionsprozessen nimmt drastisch zu, die Überwachung der Prozess wird

immer kostenintensiver und zeitaufwendiger. Aufgrund der zahlreichen Qualitäts-, Sensor- und Anlagendaten, die in der Produktion aufgezeichnet werden, entsteht großes Potenziale bezüglich der Prozessoptimierung, Qualitätsüberwachung und Instandhaltung [BERN11].

3.2 Bedeutsame KDD-Vorgehensmodellen und ihre Strukturen

Wie bei jeglichen anderen Aktivitäten, sind für das Knowledge Discovery in Databases verschiedene Modelle entwickelt und publiziert worden, die das Prozessvorgehen vereinfacht beschreiben. Für die systematische Vorgehensweise der Wissensgewinnung in Datenbeständen ist die Betrachtung unterschiedlicher Prozessmodelle relevant, wovon die wichtigsten und bekanntesten in den folgenden Abschnitten näher erläutert werden. Im wissenschaftlichen Umfeld gehören das Stufenmodell nach Fayyad et al. (1996) sowie im industriellen Umfeld das Prozessmodell *Cross Industry Standard Process for Data Mining* (CRISP-DM) nach Chapman et al. (2000) zu den weit verbreiteten Vorgehensmodellen. Die gängigen Vorgehensmodelle, abgesehen von der Variation der Anzahl, Fokus und Bedeutung der Phasen, lassen sich auf vier Kernprozessschritte reduzieren: Datenvorbereitung, Datenvorverarbeitung, Methodenanwendung und Interpretation [SHAR13].

Das KDD-Stufenmodell nach Fayyad/Piatetsky-Shapiro/Smyth

In der Literatur häufig vertretene KDD-Stufenmodell wurde von Fayyad, Piatetsky-Shapiro und Smith im Jahr 1996 entwickelt und veröffentlicht. Das Ziel des Vorgehensmodelles ist es, hochwertiges Wissen, unter Anwendung von verschiedenen Methoden, zu extrahieren. Die **Abbildung 12** gibt einen Überblick über die einzelnen Phasen des Stufenmodells im KDD-Prozess.

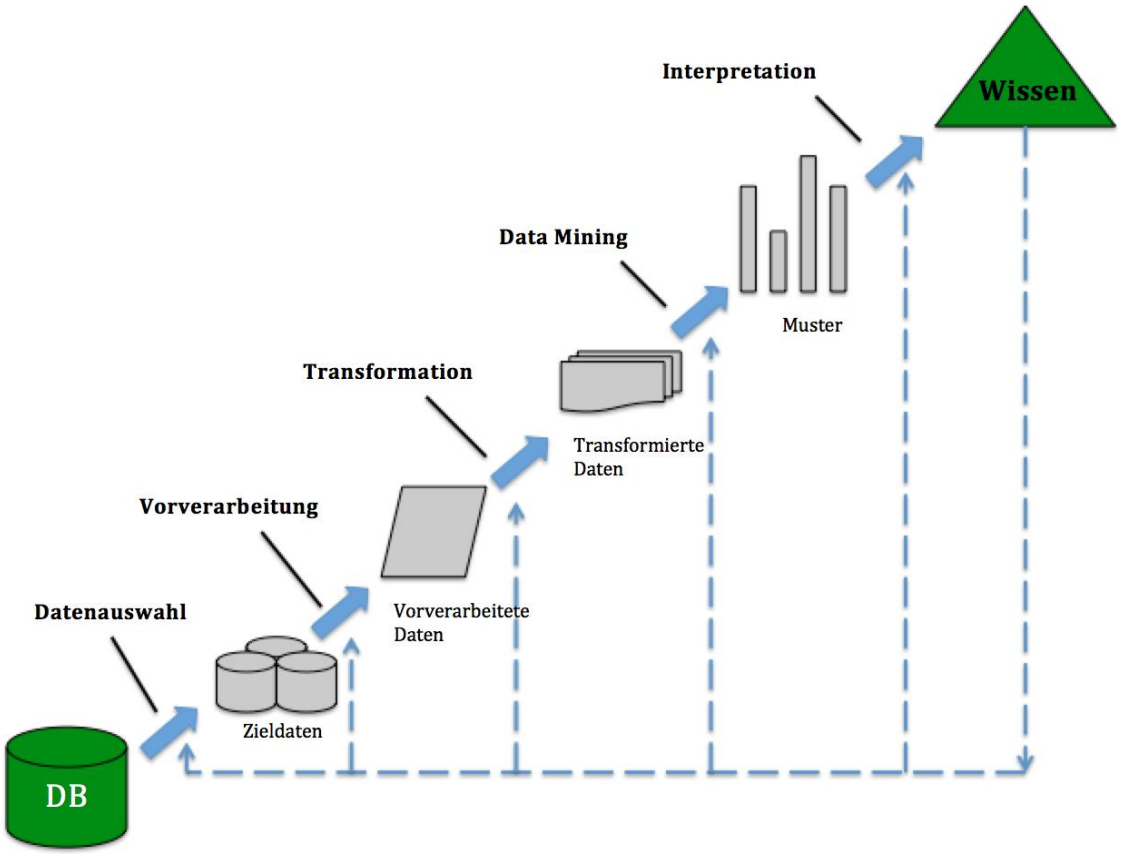


Abbildung 12: KDD-Stufenmodell nach Fayyad et al. [FAYY96]

Im Folgenden werden die zu durchlaufenden Teilprozesse detaillierter beschrieben [SHAR13].

- **Domänenverständnis und Zieldefinition:** Zu Beginn des Prozesses wird vom Anwender verlangt, sich ein Domänenverständnis anzueignen und die Ziele der Analyse festzulegen. Je nach Ziel, welches entweder die Vorhersage von tatsächlichen Aktivitäten oder die Entdeckung von Zuständen beinhaltet, wird der weitere Prozess geplant und gesteuert.

- **Datenselektion:** Abhängig vom Domänenverständnis und den Zielen, werden beim Selektionsprozess die für die Analyse benötigte Datenbasis bestimmt und aus den Datenquellen extrahiert. Um die Datengrundlage für die nachfolgenden KDD-Prozesse zu bilden, werden die Daten in einen Zieldatenbestand überführt.

- **Datenvorverarbeitung:** Die Zieldaten aus den Datenquellen beinhalten häufig Fehler und Inkonsistenzen, deshalb werden im Rahmen der Vorbereitung Datenbereinigungen durchgeführt, um die Datenqualität des Zielbestands zu gewährleisten und somit die Ergebnisse des Wissensentdeckungsprozesses nicht zu verfälschen.

- **Datentransformation:** Im Zuge der Datentransformation werden die vorverarbeiteten Daten in ein Zielformat transformiert um die nachfolgenden Analyseschritte durchführen zu können. Hierfür werden Datendimensionen reduziert und Transformationsmethoden angewendet, um die Anzahl der Variablen zu reduzieren.

- **Data-Mining:** Nachdem die Daten in den vorherigen Teilprozessen entsprechend vorbereitet wurden, beginnt im Schritt Data-Mining nun die eigentliche Datenanalyse. Um die Analyse durchführen zu können, muss für die Erreichung der Zielformulierung die passende Data-Mining Methode ermittelt werden. Das Ziel, durch den Einsatz von verschieden Analysemethoden, ist die Erkennung von Mustern und Beziehung in den transformierten Datenbeständen.

- **Interpretation und Evaluation:** Im letzten KDD-Prozess werden die mit Hilfe des Data Mining gefundenen Muster visualisiert und in Abhängigkeit mit den festgesetzten Zielen interpretiert. Das extrahierte Wissen kann direkt genutzt oder dokumentiert werden.

Das CRISP-DM Referenzmodell

Das CRISP-DM Modell von Chapman et al. hat sich bislang als Standard im industriellen Umfeld für KDD-Projekte durchgesetzt. Das CRISP-Modell dient zur Wissensentdeckung in Datenbeständen und ist anhand von praktischen Erfahrungen in DM-Projekten entwickelt worden. Wie aus der **Abbildung 13** zu entnehmen ist, lassen sich im CRISP-DM Referenzmodell sechs miteinander verknüpfte Phasen erkennen. Den Ausgangspunkt für den Prozess bilden die vorhandenen Datenbestände, die in den folgenden Phasen „Business Understanding“, „Data Understanding“, „Data Preparation“, „Modeling“, „Evaluation“ und „Deployment“ bearbeitet und ausgewertet werden. Es handelt sich hierbei nicht um einen einmaligen und sequentiellen Ablauf, vielfältige Rückkopplungen zu den vorgelagerten Phasen sind vorgesehen und erwünscht [SHAR13].

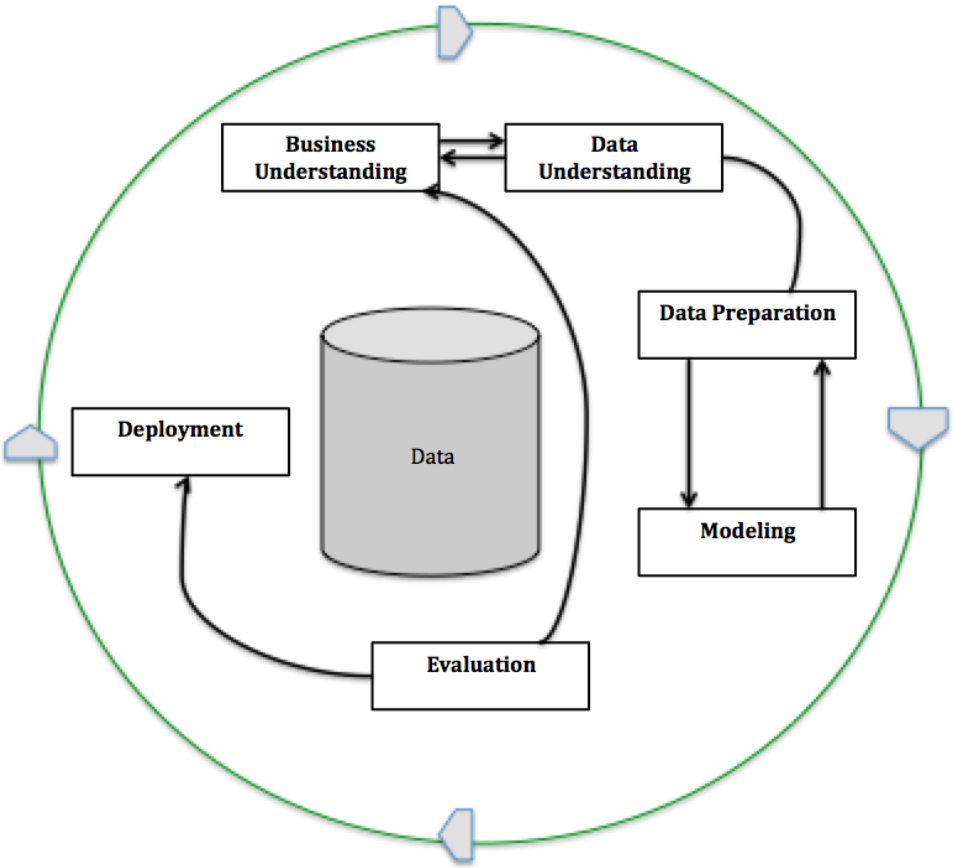


Abbildung 13: CRISP-DM Referenzmodell nach Chapman et al. [CHAP2000]

Dieses Vorgehensmodell wird hier zur Vollständigkeit und als Beispielmodell für das industrielle Umfeld genannt. Die einzelnen Phasen des CRISP-DM Modell werden in dieser Arbeit nicht ausführlich beschrieben, da im nächsten Abschnitt ein aktuelles KDD-Vorgehensmodell für die Praxis vorgestellt wird.

3.3 Das KDD-Vorgehensmodell nach MESC

Es existiert eine Vielzahl an Vorgehensmodellen mit verschiedenen Ansätzen für unterschiedliche Bereiche. Im letzten Abschnitt werden gängige KDD-Vorgehensmodelle für die Forschung und Industrie vorgestellt. Die Entwicklung dieser KDD-Modelle liegt schon mehrere Jahre zurück, dauerhaft wird nach neuen Lösungen für das aktuelle Situationsumfeld gesucht und die KDD-Modelle entsprechend angepasst. An der TU Dortmund sind an den Lehrstühlen APS und ITPL zwei unterschiedliche KDD-Modelle entwickelt worden. Speziell für das industrielle Umfeld ist das *Knowledge Discovery in Industrial Database* (kurz: KDID, s. **Anhang 1**) , als Ersatz für das CRISP-DM Referenzmodell, vom Lehrstuhl APS entwickelt worden [DEUS13]. Das *Vorgehensmodell zur Musterextraktion in Supply Chains* (kurz: MESC) ist ein KDD-Modell, dass

am ITPL entwickelt worden ist. Der Unterschied besteht darin, dass KDID ein allgemeines Modell für die Industrie darstellt und MESC sich speziell auf das Teilgebiet Supply Chain spezialisiert hat. Die Arbeit basiert auf Prozess- und Qualitätsdaten, die der Produktionslogistik zugeordnet werden. Für die Untersuchung von produktionslogistischen Daten fällt die Entscheidung auf das MESC-Vorgehensmodell und bestimmt somit die Struktur dieser Masterarbeit. Im Folgenden werden die sieben Phasen und die enthaltenen Schritte ausführlich vorgestellt [SCHE17]:

(1) Aufgabendefinition:

In der Praxis benötigt die KDD-Analyse vor der Durchführung eine konkrete Fragestellung zur Wissensentdeckung. Dafür wird in der ersten Phase die *Aufgabenstellung* (Schritt 1.1) des KDD-Prozesses bestimmt. Die Aufgabenstellung wird unter Berücksichtigung von Randbedingungen und Zielkriterien festgelegt. Die Randbedingungen können zeitliche, technische und fachliche Kategorien beinhalten.

(2) Auswahl der relevanten Datenbestände

Diese Phase gliedert sich in den Schritten der *Datenbeschaffung* (2.1) sowie anschließender *Datenauswahl* (2.2). Im Bereich der Supply-Chain liegen Informationen selten gemeinsam auf einem abrufbaren System. Die Daten liegen verteilt auf komplexen Datenquellen und können nicht nach Bedarf abgerufen werden. Die Aufgabe besteht nun darin, die richtigen Quellen zu finden und sich den Zugang zu den Zieldaten zu verschaffen. Für die Identifikation der richtigen Quellen ist der Austausch mit mehreren Projektbeteiligten nötig, da eine Einzelperson nicht das Kontextwissen verfügt, die komplexe Datenvernetzung im Unternehmen zu überblicken. Anschließend erfolgt die Datenauswahl, um die festgelegten Ziele zu erreichen, sollten auch die richtigen Informationen für die Analyse vorliegen. Um unnötig große Mengen an Daten zu extrahieren, erfolgt eine Reduktion der Datenmenge. Unter Einbeziehung von Kontextwissen können die relevanten Daten extrahiert werden.

(3) Datenvorverarbeitung

Die Datenvorverarbeitung ist die wichtigste und gleichzeitig zeitintensivste Phase des KDD-Prozesses. Sie dient der Verbesserung der Datenqualität, um somit relevante Ergebnisse zu erzielen. Für die Verarbeitung der Daten sind mehrere Schritte vorgesehen, die je nach Beschaffenheit der Datensätze angewendet werden können. Die vier Bearbeitungsschritte sind wie folgt einzuteilen: *Formatstandardisierung* (3.1), *Gruppierung* (3.2), *Datenanreicherung* (3.3) sowie die abschließende *Transformation* (3.4). Zur Durchführung der Datenanalyse benötigt das Data-Mining ein Standarddatenformat mit einer Tabelle, bei dem die Spalten, die Attribute und Zeilen die

Datensätze bilden. Mithilfe der Formatstandardisierung erfolgt die Überführung und Verknüpfung verschiedener Datenbestände anhand fester Merkmale in einen Datenbestand, womit die weiteren Schritte durchgeführt werden können. Anhand der Gruppierung erfolgt die fachliche Einteilung der Datenbestände unter Berücksichtigung der Aufgabenstellung. Falls die vorliegenden Daten nicht zufriedenstellend sind und Lücken aufweisen, kann die Datenanreicherung mittels Kontextwissen erfolgen, um neue Attribute zu erzeugen bzw. Zeilen zu füllen. Der letzte Schritt der Datenvorverarbeitung, die Transformation, dient zur Beseitigung von fehlerhaften Attributen, zur Reduzierung von Attributen und Ausreißern.

(4) Vorbereitung des Data-Mining-Verfahrens

Nach Abschluss der Datenvorverarbeitung werden in dieser Phase die Vorbereitungen für das bevorstehende Data-Mining-Verfahren getroffen. Die Phase beinhaltet die Verfahrens-, und Werkzeugauswahl sowie die fachliche und technische Kodierung. Die wichtigste Entscheidung wird in Betracht der späteren Auswertung und Visualisierung in der *Verfahrensauswahl* (4.1) getroffen. An einer Vielzahl an Data-Mining-Verfahren wie z.B. Assoziationsanalyse, Clusteranalyse, Entscheidungsbaum usw. wird das ideale Verfahren zum Projekt, definierten Aufgabe und dazugehörigen Randbedingungen ausgewählt. Nach Auswahl eines geeigneten Verfahrens muss nun im nächsten Schritt Werkzeugauswahl (4.2), eine Entscheidung über die Data-Mining Software getroffen werden. Zur Unterstützung des Data-Mining werden verschiedene Software wie RapidMiner, SPSS oder KNIME angeboten. Für die Auswahl sind jedoch Kriterien wie Datenschutz, Anpassungsfähigkeit und Systemabhängigkeit zu beachten. Falls die ursprüngliche Kodierung für das ausgewählte Verfahren nicht geeignet ist, werden die Attribute einer fachlichen (4.3) und technischen Kodierung (4.4) unterzogen. Die fachliche Kodierung beschreibt mithilfe von Kontextwissen die Attributumwandlung.

(5) Anwendung des Data-Mining-Verfahren

Nach der kompletten Vorbereitung wird in dieser Phase, das eigentliche Data-Mining auf den Datenbestand angewendet. Um die Qualität der Analyse zu erhöhen, können Verfahren mit unterschiedlichen Algorithmen angewendet werden um die Eignung zu überprüfen. Dazu sind folgende Schritte notwendig: Entwicklung des *Data-Mining-Modells* (5.1) und *Training des Data-Mining-Modells* (5.2). Für die Bewertung des DMM werden die Daten in Trainings-, Validierungs- und Testdaten unterteilt. Das Ergebnis dieser Anwendung ist das Data-Mining-Model (kurz: DMM), welches anhand der Testdaten entwickelt worden ist. Zum Schluss wird das Modell mit den Validierungsdaten auf seine Zuverlässigkeit überprüft.

(6) Weiterverarbeitung der Data-Mining-Ergebnisse

In dieser Phase werden die Ergebnisse des Data-Mining-Verfahren weiterverarbeitet und dabei relevante Ergebnisse extrahiert. Die aufbereiteten Ergebnisse werden als Wissen in das Unternehmen eingeführt. Um an das Wissen heranzukommen, wird der Schritt der *Extraktion handlungsrelevanter Data-Mining-Ergebnisse* (6.1) ausgeführt. Hierbei werden relevante Muster, unter Berücksichtigung der Handlungsrelevanz, ausgesucht und besonders betrachtet. Die ausgewählten Muster werden anhand des folgenden Schrittes *Darstellungsformation der Data-Mining-Ergebnisse* (6.2) in das Zielformat überführt.

(7) Bewertung des Data-Mining-Prozesses

Zum Abschluss des MESC-Vorgehensmodells erfolgt in der letztens Phase eine *Qualitätskontrolle des Data-Mining-Prozesses* (7.1) und die *Rückführung von Data-Mining-Ergebnissen* (7.2). Mit dem letzten Schritt wird sichergestellt, dass im Unternehmen jedem Teilnehmer das gewonnene Wissen zur Verfügung steht.

3.4 Datenvorverarbeitung

In der Praxis werden Daten, die zur Verfügung stehen, nicht immer in direkt bearbeitbarem Zustand vorgefunden. Die zunehmend voranschreitende Digitalisierung erzeugt hohe Mengen an Daten, somit steigt die Komplexität und Dimensionalität an. Bevor Data-Mining-Verfahren zum Einsatz kommen, müssen Rohdaten entsprechend vorbereitet werden. Dieser Prozess beinhaltet die Beseitigung der Daten von Fehlern, Ausreißern und Redundanz sowie die Standardisierung und Zusammenfassung. Die Datenvorverarbeitung spielt in jedem Vorgehensmodell eine bedeutsame Rolle und ist deshalb sehr arbeits- und zeitintensiv. Die Datenvorverarbeitung hat das Ziel, die Qualität der Daten zu verbessern, um somit das Laufzeitverhalten des Data-Mining Prozesses sowie die Chance auf eine erfolgreiche Datenanalyse zu erhöhen. Aufgrund der hohen Bedeutung der Datenvorverarbeitung, werden die einzelnen Verfahren, die sich dabei in *verfahrensunabhängige* sowie *verfahrensabhängige Methoden* unterscheiden, hier näher erläutert. Es wird darauf hingewiesen, dass diese Schritte als ein allgemeines Vorgehen der Datenvorverarbeitung vor einer Datenanalyse, unabhängig vom Vorgehensmodell, angewendet werden können [RUNK10]; [SHAR13].

3.4.1 Verfahrensunabhängige Methoden

Verfahrensunabhängige Methoden werden unabhängig von dem eingesetzten Data-Mining-Verfahren angewendet. Deshalb können diese vorverarbeitenden Schritte unabhängig von der

späteren Verfahrensauswahl angewandt werden [PETE05]. In diesem Abschnitt sollen die Datenvorverarbeitungsschritte und seine Methoden vorgestellt werden.

Datenintegration

Nach Auswahl der Daten stellt die Datenintegration einen wichtigen Schritt für die bevorstehende Datenanalyse dar. Die Datenmengen kommen aus verschiedenen Datenquellen und IT-Systemen und müssen schließlich zu einem Datensatz zusammengeführt werden. Zum Beispiel werden Fertigungsparameter und Logistik-Daten aus unterschiedlichen Systemen extrahiert und kombiniert. Um verschiedene Datensätze zu kombinieren, werden Merkmalsvektoren hinzugezogen. Anhand der Merkmalsvektoren werden die verschiedenen Datensätze gesucht und zugeordnet. Die Zuordnung erfolgt auf Basis von Marken, dies können z.B. Codes, Zeitstempel, Seriennummer oder Ortsangaben sein. Die markenbasierte Zuordnung wird anhand der **Abbildung 14** verdeutlicht. Falls keine eindeutige Zuordnung erfolgt, führt dies zu fehlenden Einträgen und werden entfernt [RUNK10].

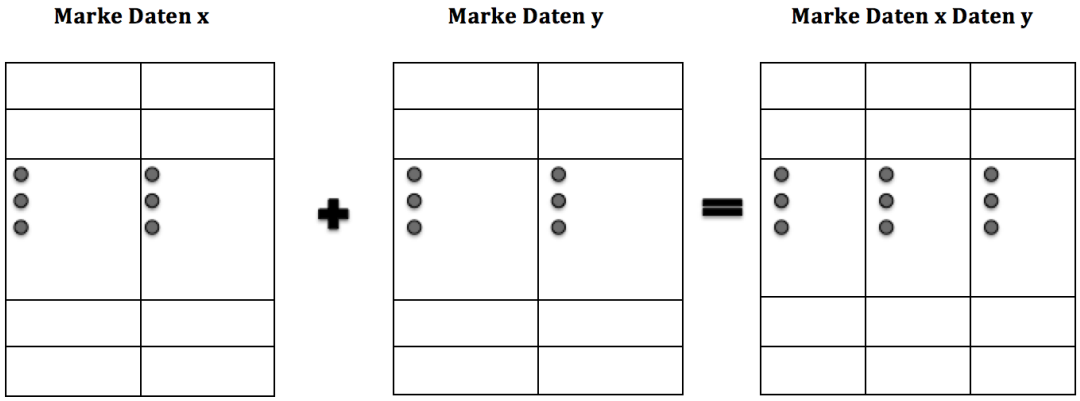


Abbildung 14: Markenbasierte Integration von Datensätzen nach [RUNK10]

Datenbereinigung

Mehrere Datensätze, die anhand der Marke zusammengeführt worden sind, beinhalten nach der Kombination nun eine größere Menge an Daten. Je größer Datensätze werden, desto mehr Fehler können diese aufweisen. Um eine effiziente Datenanalyse durchführen zu können, ist die Bereinigung von fehlenden und verrauschten Daten sowie das Aufspüren von Ausreißern von großer Bedeutung. Die Bereinigung sollte neutral erfolgen, zusätzliche Informationen dürfen nicht eingefügt werden, um die vorhandenen Informationen nicht zu verzerren oder zu verfälschen. Im Folgenden werden die häufigsten Probleme vorgestellt.

- **Fehlende Daten**
  In seltenen Fällen liegen Daten vollständig und korrekt vor, jedoch ist dies eine Voraussetzung für die genaue DM-Analyse. Um die fehlenden Daten zu kompensieren,

bieten sich zwei gängige Lösungen für dieses Problem an. Die fehlenden Werte werden durch den häufigsten Wert bzw. Mittelwert ersetzt oder das komplette Attribut aus der Daten-Tabelle gestrichen [CLEV16].

- **Verrauschte Daten**
  Unter verrauschte Daten werden fehlerbehafte Daten verstanden, die durch zusätzliche Einflüsse verfälscht worden sind und somit Schwankungen aufweisen. Um diese Einflüsse auf dem Datensatz zu verringern, werden die Daten geglättet. Das Glättungsverfahren ist ein spezieller Filter, wobei der zu bearbeitende Datenpunkt mit einigen Nachbarwerte verglichen und daraus der korrigierte Datenpunkt berechnet wird [CLEV16].

- **Ausreißer**
  Ausreißer sind dadurch gekennzeichnet, dass sie in der Gesamtmenge selten auftauchen und von der Verteilung der übrigen Daten drastisch abweichen. Um potentielle Ausreißer zu erkennen bieten sich Methoden, wie die Verbundbildung (Clustering) an. Abweichungen liegen außerhalb des Cluster und können wie fehlende Daten behandelt werden [CLEV16].

Datenreduktion

Aufgrund der hohen Datenmengen, die Data-Mining ausgesetzt sind, ergeben sich Leistungsschwankungen, die schließlich die Datenanalyse erschweren. Um diese Problematik zu vermeiden, können Datenmengen reduziert werden. Die Reduktion erfolgt anhand der Verringerung der Anzahl der Attribute (Aggregation und Dimensionsreduktion) bzw. Anzahl der Datensätze (Stichprobenziehung).

- **Aggregation**
  Die Aggregation fasst die Datensätze untergeordneter Ebene zu einem Datensatz höherer Aggregationsebene zusammen. Beispielsweise können *Tag* und *Woche* zu der höheren Aggregationsebene *Monat* zusammengefasst werden. Dieses Verfahren verringert somit das Datenvolumen, jedoch ist zu beachten, dass damit auch ein Informationsverlust verbunden ist [PETE05].

- **Dimensionsreduktion**
  Anhand der Dimensionsreduktion wird versucht einen hoch dimensionalen Datensatz in einen niedrigen dimensionalen Datenbestand zu überführen, ohne dabei Informationen für die weitere Data-Mining Analyse zu verlieren. Die Reduzierung des Datenbestandes

erfolgt durch das Entfernen irrelevanter oder redundanter Attribute. Die irrelevanten Attribute sind durch eine schwache Korrelation und redundante Attribute durch eine hohe Korrelation mit Klassifikationsattribute erkennbar [PETE05].

- **Stichproben**
  Die Durchführung der DM-Analyse auf den kompletten Datensatz führt zu Leistungseinschränkungen, weshalb eine Reduktion der Daten von Vorteil sein kann. Für das Ziehen der Stichprobe ist zu beachten, dass die Stichprobe einen bestimmten Anteil einer Gesamtmenge ausmacht und somit den realen Zusammenhang der Gesamtheit wiederspiegeln muss. Die Stichprobe kann rein zufällig oder durch vorher fest gelegte Kriterien entnommen werden [PETE05].

3.4.2 Verfahrensabhängige Methoden

Verfahrensabhängige Methoden kommen erst zur Anwendung, sobald das bevorstehende Data-Mining Verfahren konkret festliegt. Somit gehört dieser Schritt nicht direkt zur klassischen Datenvorverarbeitung, die unabhängig vom Verfahren durchgeführt werden kann, wird aber zum besseren Verständnis in diesem Abschnitt kurz erläutert [CLEV16]. Im MESC gehört dieser Schritt eher zur *fachlichen* (Schritt 4.3) und *technischen Kodierung* (Schritt 4.4) in der Phase *Vorbereitung des Data-Mining-Verfahren.*

Datentransformation

Trotz der klassischen Vorbereitung der Daten, ist die ursprüngliche Form noch nicht ausreichend für Data-Mining geeignet. Die Transformation hat nun die Aufgabe, die Daten in die erforderliche Form zu überführen und somit an das entsprechende DM-Verfahren anzupassen, mit der es auch arbeiten kann. Für die Überführung existieren verschiedene Transformationsmethoden, die hier erläutert werden.

- **Anpassung des Datentyps**
  Um verschiedene Analysemethoden anwenden zu können, sind verschiede Datenformen nötig. Es existieren eine Vielzahl unterschiedlicher Datenformen, jedoch unterscheidet man zwischen drei relevanten Datentypen: nominal, ordinal und metrische Daten [HATZ09]. Nominale Daten liegen rein in qualitativer Merkmalsausprägung (z.B. Geschlecht, Beruf) vor und unterliegen keiner Rangfolge. Sie lassen sich nur durch *gleich* oder *ungleich* abgrenzen und somit sind nur Angaben über Häufigkeiten und Anteile machbar. Ordinale Daten sind den nominalen Daten sehr ähnlich, nur hier existiert eine natürliche und feste Rangfolge (z.B. sehr gut, gut, mittel, schlecht). Metrische Daten bestehen aus Zahlenwerten (z.B. Körpergröße, Anzahl Schüler, Dauer eines Vorgangs),

somit sind die Voraussetzungen gegeben, mathematische Operationen anzuwenden [CLEV16].

- **Diskretisierung (Binning)**
  Dieses Verfahren dient dazu, die Anzahl der Werte eines numerischen Attributs zu reduzieren. Dabei wird der Wertebereich der Attribute, der Größe nach aufsteigend, in sogenannte *Bins* eingeteilt. Die Einteilung in *Bins* sorgt dafür, dass die Granularität der Daten reduziert wird [CLEV16].

- **Kombination oder Separierung**
  Durch Zusammenfügen verschiedener Attribute zu einem neuen Attribut oder die Zerlegung eines Attributs in seine einzelnen Bestandteile, kann je nach Verfahren notwendig sein, um weitere neue Informationen zu gewinnen. Ein Beispiel wäre das Zusammenfassen von Tag, Monat, Jahr zu Datum. Die Umkehrung wäre die Zerlegung vom Datum in seine Bestandteile [CLEV16].

3.5 Data-Mining-Verfahren

Das Data Mining-Verfahren ist die anspruchsvollste und entscheidendste Phase des KDD-Prozesses, um Wissen aus Datenbeständen zu extrahieren. Um Trends oder Muster zu erkennen, umfasst dieser Teilschritt diverse Algorithmen und Methoden, die zum Einsatz kommen. Aufgrund der Relevanz dieses Teilschrittes für den weiteren Verlauf der Arbeit, wird der Begriff „Data Mining“ und seine Analysemethoden näher erläutert.

3.5.1 Definition und Beschreibung von Data-Mining

Wie bereits im letzten Abschnitt erwähnt, handelt es sich beim Data Mining um einen Teilschritt des KDD-Prozesses. Der Begriff „Data Mining“ reicht bis in die 1980er Jahre und lässt sich buchstäblich mit „Graben, Schürfen in Daten oder Datenabbau“ übersetzen [KEUP09]; [PETE05]. Cleve definiert Data-Mining als „Extraktion von Wissen aus Daten“ [CLEV16]. Der Prozess besteht aus bestimmten Algorithmen, die durch die Anwendung auf bestimmte Datensätze Auffälligkeiten und bestimme Muster erkennen. Bei der konkreten Algorithmen Entwicklung sind hauptsächlich Forschungsbereiche aus der Mathematik oder Statistik beteiligt. In verschiedenen Literaturen werden KDD und Data-Mining gleichbedeutend für den gesamten Prozess verwendet. Fayyad et al. (1996), gelten als Wegbereiter des Data-Mining, sehen KDD als gesamten Prozess an, die für die Wissensentdeckung aus Datenmengen zuständig ist. Das Data-Mining wird als ein Teilprozess des Gesamtprozesses der Wissensentdeckung angesehen, wobei die Datenvorverarbeitung nicht mit einbezogen wird. Hierbei ist anzumerken, dass die Basis für

eine erfolgreiche DM-Analyse, erst durch die Datenvorverarbeitung ermöglicht wird. Sonst besteht die Gefahr, dass entdeckte Wissen und Muster ungültig sind und somit keine Relevanz für den untersuchten Bereich haben [SHAR13].

3.5.2 Methoden des Data-Mining

Nach der Datenvorverarbeitung sind die Daten soweit aufbereitet, dass die Algorithmen angewendet werden können. Anhand dieser Algorithmen und Methoden werden schließlich wichtige Erkenntnisse und Muster identifiziert und tragen zu Wissensentdeckung bei. Dieser Abschnitt stellt Verfahren des Data-Mining vor, dazu werden die wichtigsten und gängigsten Verfahren wie Klassifikation, Clusteranalyse sowie Assoziationsanalyse vorgestellt.

Klassifikation

Die Klassifikation gehört zu der Kategorie des *überwachten Lernens* und ist die geläufigste Methode des Data-Mining. Mit dem *überwachten Lernen* ist die Fähigkeit eines Systems gemeint, Gesetzmäßigkeiten nachzubilden und möglichst zielsichere Voraussagen zu treffen. Für jedes Input x wird - sofern vorhanden- das passende Output y zugeordnet und für Trainingszwecke genutzt, um das Modell zu verfeinern und zu optimieren. Bei dieser Methode werden Objekte anhand bestimmter Merkmale einer Klasse zugeordnet. Die Zuordnung erfolgt über einen Algorithmus, auch Klassifikator genannt. Dabei wird das Ziel verfolgt, für neue Objekte eine präzise Vorhersage einer Klassenzugehörigkeit zu ermöglichen [CLEV16].

- **Entscheidungsbaum**
  Die Erhebung von Wissen und Informationen dienen dazu, präzise Entscheidungen zu treffen, dabei muss auch ersichtlich sein, wie diese zustande kommen. Eine gute Darstellungsform von hierarchisch aufeinander folgende Entscheidungen bietet ein *Entscheidungsbaum (s.* **Abbildung 15***).* Der Entscheidungsbaum ist eine Methode zur automatisierten Klassifikation von Datenobjekten. Hierbei ist der Weg zur Entscheidung, durch die graphische Darstellung sowie mit der zusätzlichen Begründung, leicht nachvollziehbar. Die Entscheidungsbaumdiagramme beginnen mit einem einzelnen *Wurzelknoten* und verzweigen sich durch die darauffolgenden Entscheidungsmöglichkeiten nach unten. Die viele Verzweigungen bilden ein *Wurzelnetzwerk,* wobei die *Wurzelspitze (Blattknoten)* ein fertiges Konzept mit den relevanten Entscheidungen darstellen [DREW10]; [CLEV16].

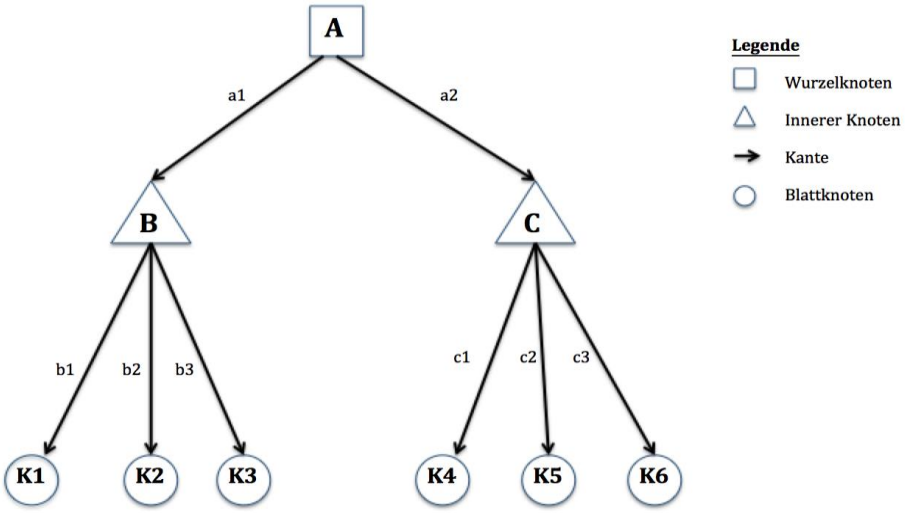


Abbildung 15: Darstellung eines Entscheidungsbaumes nach [DREW10]

ID3-Algorithmus

Zu dem bekanntesten Algorithmus des Entscheidungsbaumes gehört der ID3-Algorithmus und dient als Grundlage auf den aufbauenden Algorithmen. Mit seiner rekursiven und iterativen Vorgehensweise leicht nachvollziehbar und verständlich für den Anwender. Der ID3 ist in der Lage aus großen Datenmengen Entscheidungsbäume zu generieren und bietet eine hohe Klassifikationsgenauigkeit. Die Klassifizierung wird anhand einer Minimalkombination von Attributwerte aufgebaut, hierzu genügen wenige Attribute aus einer Vielzahl an Attributen. Ein ID3-Entscheidungsbaum besteht aus drei Bestandteilen: Knoten, Kanten und Blatt. Die Attribute bilden den Knoten eines Baumes, die Kanten enthalten die Attributwerte und die Blätter charakterisieren die Klassen.

Der Aufbau eines Entscheidungsbaumes beginnt mit der Auswahl eines Attributes für den obersten Knoten, dies geschieht unter Berücksichtigung von definierten Kriterien. Anhand des Top-Down-Ansatzes wird der Baum sukzessiv Knoten für Knoten nach unten ausgebaut. Die Auswahl von Attributen für den Knoten erfolgt hierarchisch, das Attribut mit dem höchsten Informationsgewinn für die jeweilige Stelle wird zuerst genommen [BEHN08].

Clusteranalyse

Unter Clustering wird die „automatische Gruppierung von ähnlichen Objekten“ verstanden [SHAR13]. Es gehört zum *unüberwachten Lernen* und lässt sich auch als Klassifizierung oder

Segmentierung bezeichnen. Beim *unüberwachten Lernen* konzentriert sich der Algorithmus darauf, Regeln für Inputdaten zu bilden. Hierbei sind nur die Inputdaten x vorhanden, ohne bekannte Zielwerte. Der Algorithmus versucht, anhand dieser Inputdaten, Muster zu erkennen [CLEV16]. Der Unterschied zwischen überwachten und unüberwachten Lernens besteht darin, das ein Modell im überwachten Lernen anhand von Daten trainiert und auf Testdaten angewendet wird, um eine präzise Vorhersage zu machen [CRON10]. Die Clusteranalyse befasst sich mit der Zusammenfassung von Datenobjekten zu Clustern oder Gruppen, hierfür wird ein Ähnlichkeitskonzept erstellt. Die Daten werden dabei partitioniert und beinhalten Datenobjekte, die sich hinsichtlich ihrer Merkmalsausprägung durch eine hohe Homogenität auszeichnen. Während Datensätze innerhalb eines Segmentes homogene Merkmale prägen, besitzen sie im Vergleich zu anderen Segmenten heterogene Ausprägungen. Die Grundlage für die Segmentbildung bildet das Abstandsmaß, womit sich die Ähnlichkeit von Datenobjekten bestimmen lässt. Die Clusteranalyse verfügt über ein breites Spektrum an Algorithmen, die zur Klassifizierung von Datensätzen angewendet werden. Laut Chamoni et al. wird zwischen *partitionierenden* und *hierarchischen Verfahren* unterschieden [CHAM06]; [SHAR13].

- **Partitionierende Clusterverfahren**

  Im Rahmen des *partitionierenden Verfahren*, ausgehend von einer vorgegebenen Gruppeneinteilung (Anfangspartition), erfolgt eine Verlagerung der Objekte zwischen den Clustern, um die vorhandene Gruppenteilung zu optimieren. Dieses Verfahren erfolgt solange, bis sich die Güte der Partitionierung sich nicht mehr ändert (s. **Abbildung 16**) [CHAM06]; [CLEV16].

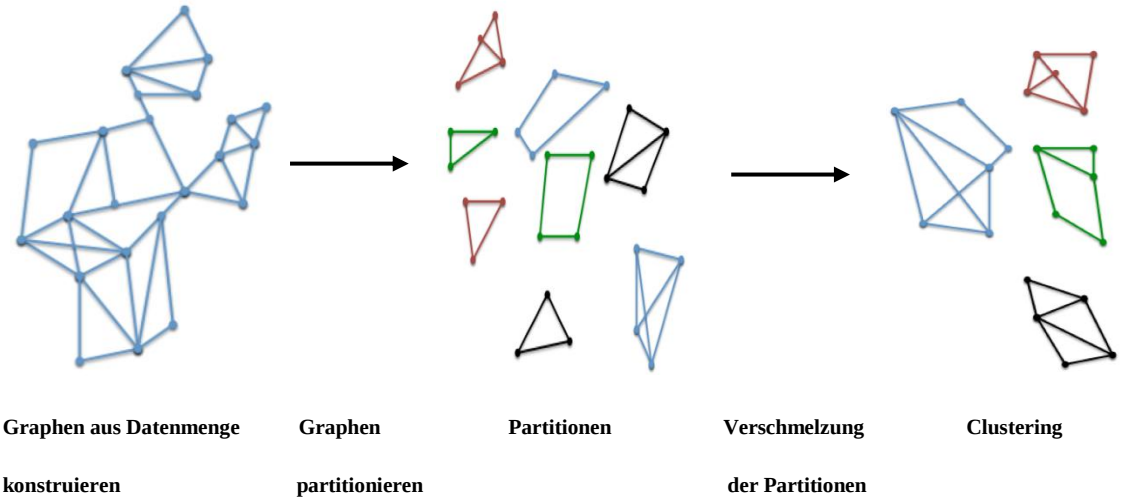
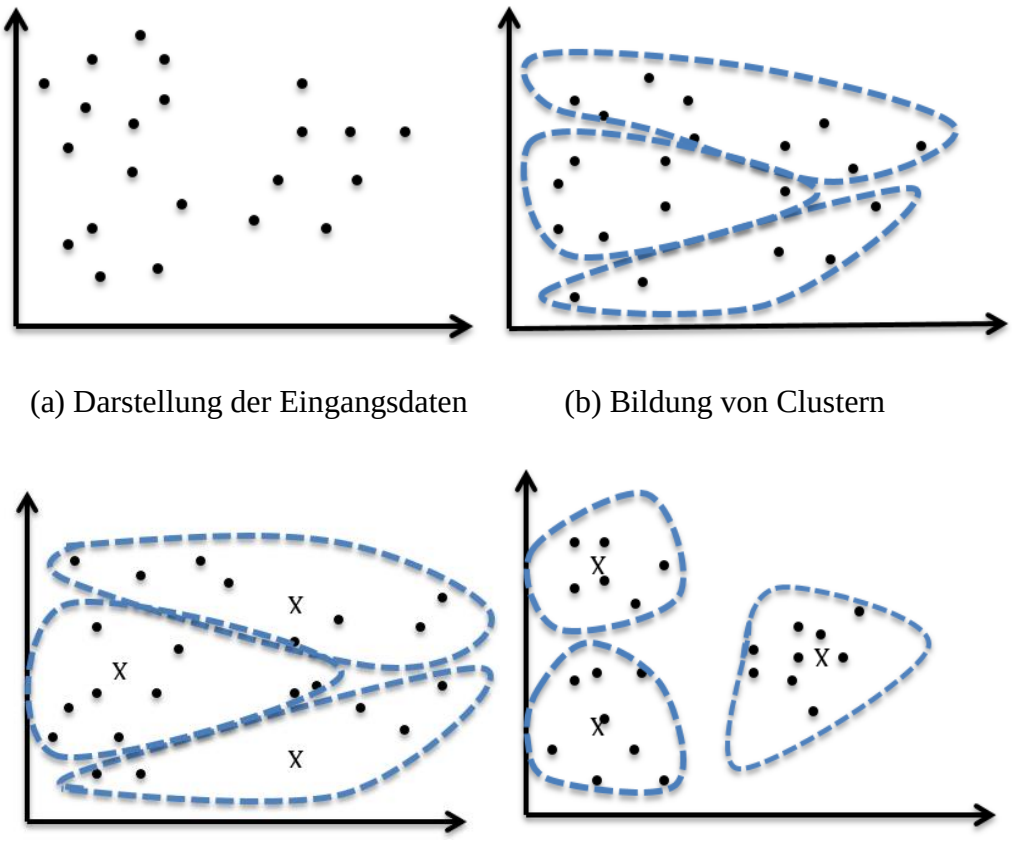


Abbildung 16: Partitionierendes Clusterverfahren nach [CLEV16]

k-Means-Algorithmus

Data-Mining-Verfahren bestehen allgemeinen aus Algorithmen, die eine eindeutige Handlungsvorschrift zum Lösen von Problemen vorgeben. Zu einem Data-Mining-Verfahren wie z.B. dem Cluster existieren zahlreiche Algorithmen, die sich vom Aufbau und Ausführung unterscheiden. Der k-Means-Algorithmus gehört zu der bekanntesten und häufigsten Handlungsvorschrift der partitionierenden Clusteranalyse. Dieses Algorithmus verfolgt das Ziel, den Datensatz in die vorgegebene k-Anzahl an Clustern zu unterteilen (k= Anzahl der Cluster). Die Variable k steht für die Anzahl der Cluster und kann vorgegeben werden. Die Bildung der Cluster erfolgt erst initial durch einen Zufallsgenerator, damit wird eine grobe Struktur für den Clusteraufbau geschaffen. Mithilfe dieser Basis werden die Centroide berechnet, die als Anlaufzentrum dienen und die umliegenden Punkte in einer erneuten Zuordnung mit der geringsten Distanz zuordnen [CLEV16]. In **Abbildung 17** ist der Ablauf des k-Means-Algorithmus zum Verständnis dargestellt.



(a) Darstellung der Eingangsdaten (b) Bildung von Clustern

(c) erste Auswahl der Clusterzentren (d) Partitionierung durch k-Means

Abbildung 17: Clustering mit dem k-Means-Algorithmus nach [CLEV16]

- **Hierarchisches Clusterverfahren**

  Das *hierarchische Verfahren* benötigt dagegen keine vorgegebene Gruppenteilung, sondern teilt die Datenmenge gestaffelt in Clustern ein. Innerhalb dieses Verfahren lässt es sich wiederum in *agglomerativen* und *divisiven Clusterbildung* (s. **Abbildung *18***) unterteilen.

  Die *agglomerative* Clusterbildung fängt mit der kleinsten Cluster-Auflösung an, jeder Datensatz bildet dabei ein eigenes Cluster. In mehreren Schritten werden diese „ein Element-Cluster“, mit ähnlichem Charakter, zu einem hierarchisch höheren Cluster zusammengefasst. Dieser Vorgang setzt sich fort, bis nur noch ein Cluster vorhanden ist. Die divisive Clusterbildung nutzt am Anfang die gesamte Datenmenge als ein großes Cluster und teilt sie in mehreren Schritten zu kleineren Untergruppen. Dieser Vorgang wird solange wiederholt, bis am Ende jedes Cluster ein Datensatz beinhaltet [CLEV16]; [CHAM06].

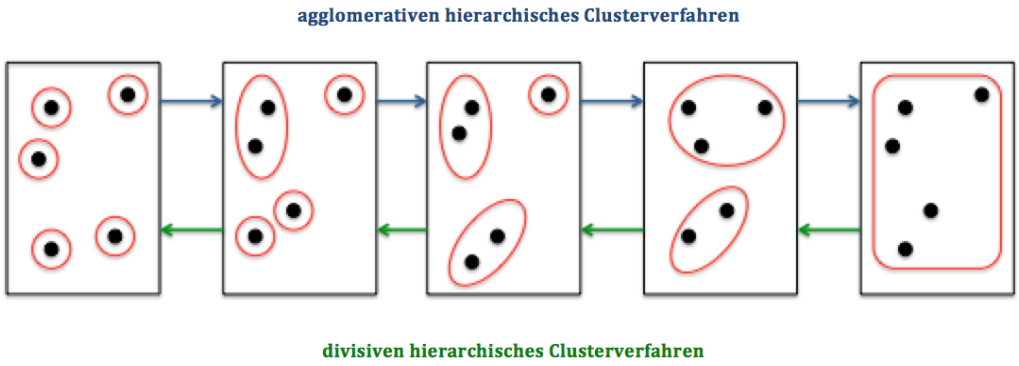


Abbildung 18: Hierarchisches Verfahren nach [CLEV16]

Assoziationsanalyse

Die Assoziationsanalyse wird zu dem überwachten Lernen zugeordnet und bezeichnet die Suche nach Regeln. Die daraus folgende Assoziationsregel beschreibt die Beziehungen zwischen sogenannten Items. Sie stellen die Elemente einer Menge bzw. Datensätze dar, die wiederum das Auftreten eines Items innerhalb einer Transaktion einbeziehen. Der Unterschied zwischen Assoziationsanalyse und Klassifikation besteht darin, dass sich nicht nur auf ein Zielattribut beschränkt wird. Es werden mehrere Beziehungen zwischen beliebigen Items offengelegt und stellt diese in Form von generierten Regeln wie z.B. „Wenn-Dann“ Beziehung dar. Eine simple Regel kann wie folgt formuliert werden: „Wenn ein Kunde das Produkt A kauft, kauft er auch Produkt B“ [CLEV16]; [PETE05]. Zur Erstellung der Regel verwendet die Analyse zwei Maßzahlen, den Support und Konfidenz. Um den Support, drückt die relative Häufigkeit aus, zu ermitteln, werden alle Items, die den festgelegten minimalen Support überschreiten, ermittelt und

in Relation zur gesamten Datenbasis gesetzt. Die Relevanz der Regel ist umso höher, je höher der Support Wert ist. Die zweite Maßzahl Konfidenz misst die Sicherheit der entdeckten Regel und spiegelt somit die Stärke für diese Regel wieder. Dabei sollte beachtet werden, dass bestimmte Mindestwerte bei beiden Maßzahlen überschritten werden müssen, um überhaupt für die weitere Generierung relevant zu sein [CLEV16]; [CHAM06].

FP-Growth-Algorithmus

Bei dem Frequent-Pattern-Growth Algorithmus (FP-Growth) handelt es sich um den Nachfolger des Apriori-Algorithmus, er gehört zu den bekanntesten Algorithmen der Assoziationsanalyse, die für die Erstellung von Assoziationsregeln gelten. Die Entwicklung des Nachfolgers erfolgte aufgrund der hohen Laufzeit durch aufwendige Datenbankdurchläufe. Um diesen Nachteil zu umgehen, werden beim FP-Growth Algorithmus die Transaktionen der Datenbank in einer Baum-Struktur, im Frequent-Pattern-Tree dargestellt. Es werden keine *candidate itemsets* erzeugt, sondern nur ein *1-elementiges item*. Die Datenbasis liegt im FP-Tree in komprimierter Form vor, aufwendige Datenbankdurchläufe werden vermieden und führen zu kürzeren Laufzeiten, denn die Daten werden als *frequent itemsets* extrahiert [CLEV16].

Der Ablauf des FP-Growth Algorithmus verfolgt das Ziel, die Datenbank in einer FP-Tree zu komprimieren und nach häufigen Mengen zu suchen. Die Funktionsweise wird hier kurz erläutert. Zunächst wird der Support, relative Häufigkeit eines Items, durch einen kompletten Datenbankdurchlauf ermittelt. Die gefundenen *1-elementiges itemsets* werden nach absteigender Häufigkeit (Supportwert) geordnet und nicht *frequente items* werden entfernt (s. **Tabelle 1**). Anhand eines zweiten Datenbankdurchlaufs wird, mithilfe der *frequenten items*, ein FP-Tree aufgebaut. Mithilfe von Header Table kann die Häufigkeit eines *items* eingesehen werden und dient als Unterstützung bei der Erstellung eines FP-Tree (s. **Abbildung 19**) [CLEV16].

Tabelle 1: 1-elementiges Item und frequent items

Transaction id	Items	frequent items
100	{ f, a, c, d, g, i, m, p, t, y}	{f, c, a, m, p)
200	{a, b, c, f, l, m, x, z}	{f, c, a, b, m}
300	{b, f, h, j, o, w}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

Item	frequency
f	4
c	4
a	3
b	3
m	3
p	3

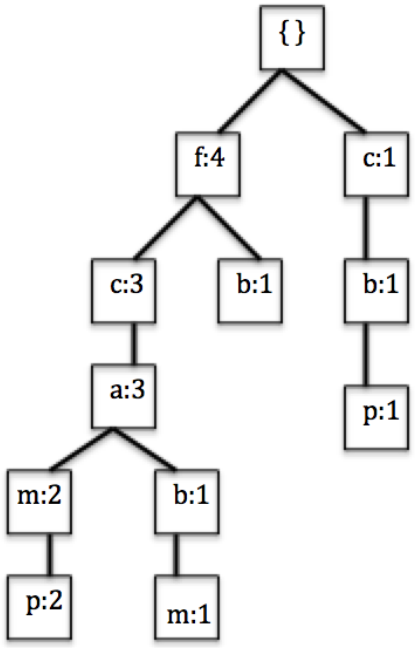


Abbildung 19: Header Table und FP-Tree

3.5.3 Auswahl und Funktionsweise des Data-Mining Software

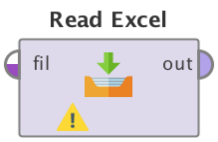
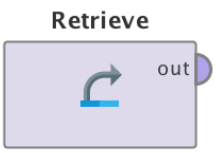
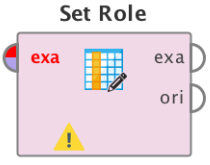
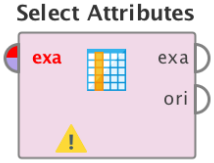
Die DM-Software, womit die DM-Analyse in produktionslogistische Daten durchgeführt wird, heißt RapidMiner. Es ist eine Umgebung für Data-Mining Prozesse und ist im Jahr 2001 am Lehrstuhl für Künstliche Intelligenz der Technischen Universität Dortmund unter dem ursprünglichen Namen YALE (**Y**et **A**nother **L**earning **E**nvironment) entwickelt worden. Dies ist ein Open Source Data-Mining Programm und kann kostenlos über die Firmenwebsite bezogen werden. Die Einsatzmöglichkeiten sind breit gefächert, die Anwendung kann sowohl in der Forschung als auch im unternehmerischen Bereich erfolgen. Hierbei ist zu erwähnen, dass es eine Vielzahl an weiteren DM-Programmen wie z.B. KNIME oder WEKA auf dem Markt zur Verfügung stehen, jedoch schneidet RapidMiner unter den Open Source Data-Mining Tool, hinsichtlich der Technologie und Anwendbarkeit, am besten ab. Die große Anzahl an Operatoren und die einfache Anwendung ohne vorherige Programmierkenntnisse, sprechen für den Einsatz von RapidMiner [Rapi10]; [SHAR13]. In dieser Arbeit wird mit der aktuellsten Version des RapidMiner gearbeitet und handelt hierbei um Version 8.0.

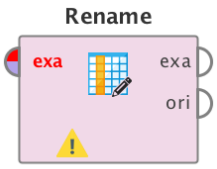
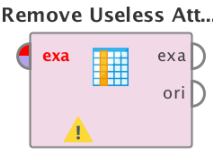
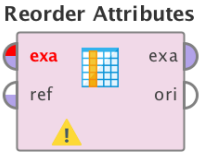
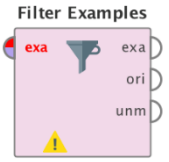
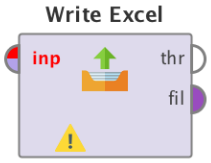
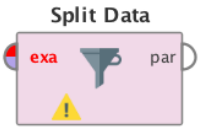
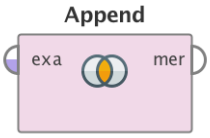
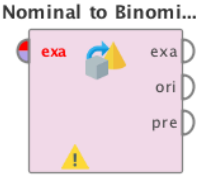
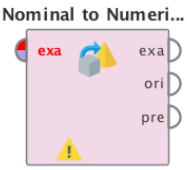
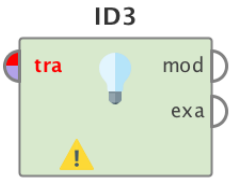
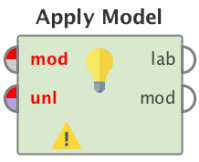
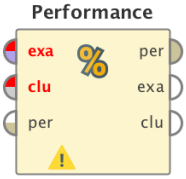
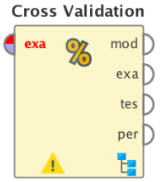
In RapidMiner werden Datenanalysevorgänge, anhand von Operatoren, als Prozess modelliert. Jeder Operator des Prozessvorgangs gibt eine gekapselte Aktivität wieder, bedient sich einem Dateninput, der verarbeitet und wieder ausgegeben wird. Für die Prozessauslegung stehen 417 Operatoren zur Verfügung, deren Funktionalität in folgenden Kategorien unterteilt werden

[Rapi10]; [SHAR13]: *Data Access, Blending, Cleansing, Modeling, Scoring, Validation, Utility und Extensions*. In **Tabelle *2*** werden Operatoren für die Prozessauslegung vorgestellt, um sich bereits ein Bild davon machen zu können. Dort werden von den 417 zur Verfügung stehenden Operatoren nur ein Bruchteil vorgestellt, aber es handelt sich hierbei um die gängigsten Operatoren, die häufig zum Einsatz kommen.

Der Aufbau des Programms wird anhand der **Abbildung 20** näher erläutert. Um einen Prozess auszulegen, müssen Daten importiert und direkt auf deren Nutzbarkeit überprüft werden. Des Weiteren besteht die Möglichkeit die Datei aus der Excel-Datei zu lesen. Falls Beanstandungen auftauchen, werden diese angezeigt. Nach erfolgreichem Datenimport befinden sich die Daten für die weitere Benutzung auf der linken Seite im *Repository* und können beliebig aufgerufen werden. Die Operatoren, um die einzelnen Prozessschritte auszulegen, befinden sich ebenfalls im linken Bereich unter dem Punkt *Operators*. Ohne jeglichen Aufwand können sie per Drag & Drop in den mittleren *Process*-Bereich eingefügt werden. Der Datenaustausch zwischen den Operatoren erfolgt durch Ports, in dem diese über Links verbunden werden und somit eine Schnittstelle zwischen den Operatoren bilden. Falls Operatoren Parameter besitzen, können diese im rechten Bereich unter *Parameters* verändert werden und somit den Prozess optimieren. Die Parameter werden mit ihren Kennzahlen angezeigt und können beliebig verändert bzw. angepasst werden. Ein weiteres zentrales Element, während des Analyseprozesses, ist der *Problems, Help & Comment view*, es wird direkt im *Process*-Feld angezeigt. Hier werden alle Fehlermeldungen und Hilfestellungen, vor bzw. während der Durchführung des Prozesses, angezeigt. Das System zeigt aufgrund der Informationensammlung des weltweiten User-Netzwerks, mögliche Vorschläge an Operatoren an, um die begonnene Prozessauslegung zu vollenden.

Tabelle 2: Beschreibung der Funktionalität von häufig eingesetzten Operatoren

Dieser Operator liest ein ExampleSet aus der angegeben Excel-Datei		Dieser Operator kann auf gespeicherte Informationen im Repository zugreifen und sie in den Prozessfeld laden	
Dieser Operator wird verwendet, um die Rolle eines oder mehrerer Attribute zu ändern bzw. festzulegen.		Dieser Operator wählt eine Teilmenge der Attribute eines Example Set aus.	

Dieser Operator kann verwendet werden, um ein oder mehrere Attribute von einem ExampleSet umzubenennen		Dieser Operator entfernt nutzlose Attribute aus einem ExampleSet. Die Grenzwerte für nutzlose Attribute werden vom Benutzer festgelegt	
Mit Operator können die regulären Attribute eines ExampleSet neu-strukturiert werden.		Dieser Operator wählt aus, welche Daten eines ExampleSets beibehalten und entfernt werden sollen.	
Dieser Operator schreibt ein ExampleSet in eine Excel-Datei.		Dieser Operator erzeugt die gewünschte Anzahl an Teilmengen und partio-niert das ExampleSet in relativen Größen.	
Dieser Operator ver-bindet zwei Example Sets mit einem oder mehreren Schlüsselattributen zu einem ExampleSet.		Dieser Operator erstellt ein fusioniertes Example Set aus zwei oder mehr kompatiblen ExampleSet, indem er alle Beispiele zu einem kombinierten Set hinzufügt.	
Dieser Operator ändert den Typ, der nicht binomialen Attribute, in einen binomialenTyp.		Dieser Operator ändert den Typ, der nicht numerischen Attribute, in einen numerischen Typ.	
Dieser Operator generiert ein Entscheidungsbaum-modell mit dem ID3-Algorithmus		Dieser Operator führt das Clustering unter Ver-wendung des k-Means-Algorithmus durch.	
Dieser Operator wendet ein Modell auf ein ExampleSet an.		Dieser Operator ist ein Algorithmus der Assoziationsanalyse. Für den Einsatz ist zu beachten, dass alle Attribute des ExampleSet binomial sind.	
Dieser Operator wird zur Leistungsbewertung ver-wendet.		Dieser Operator führt eine Kreuzvalidierung durch, um die statistische Leistung eines Modells zu schätzen.	

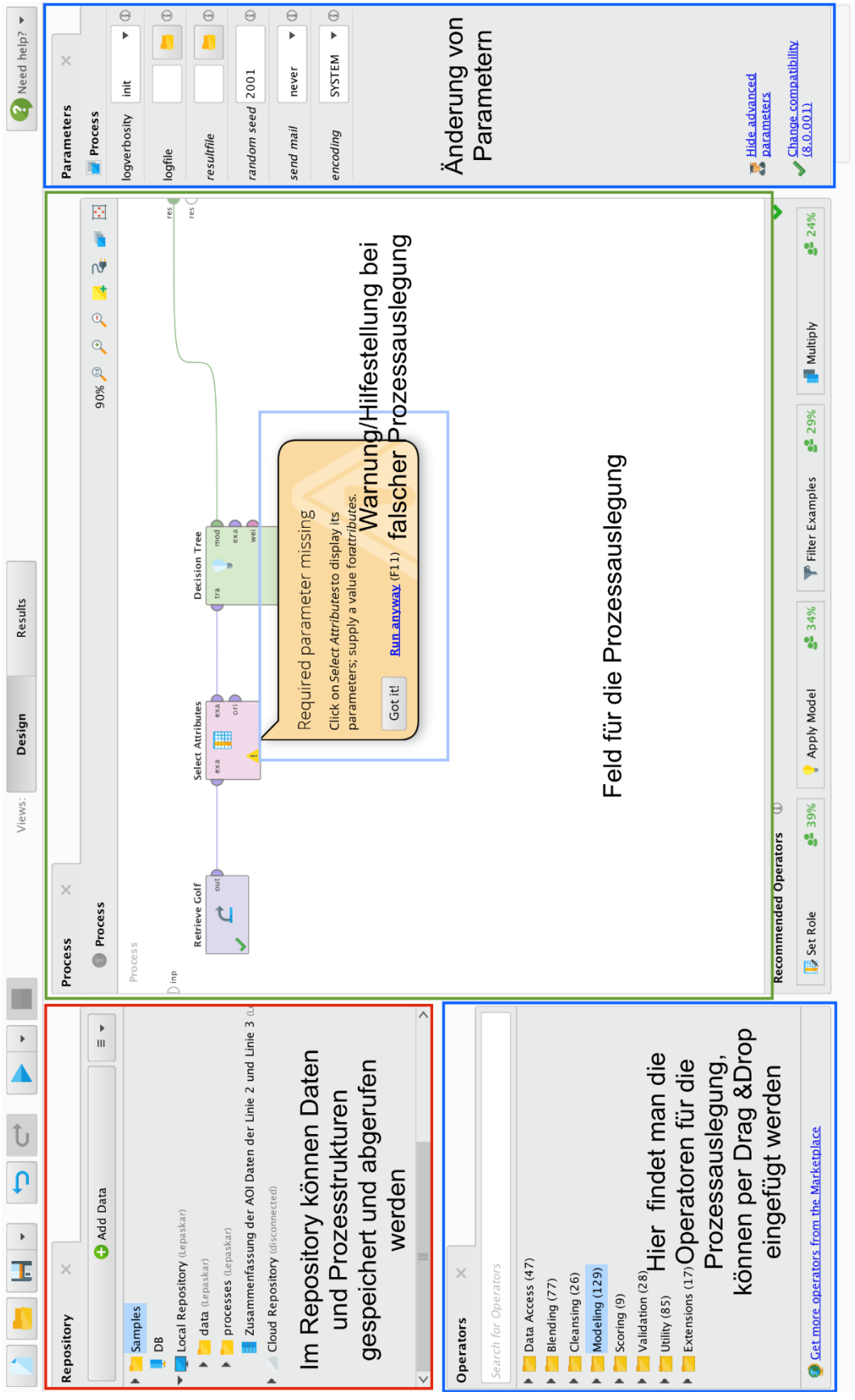


Abbildung 20: RapidMiner Programmübersicht

4. Wissensgewinnungsprozess in der Elektronik-fertigung

Im letzten Kapitel ist der Prozess der Wissensgewinnung mit seinen relevanten Vorgehensmodellen und Data-Mining-Verfahren ausführlich dargestellt und erläutert worden. Die Struktur und Aufbau des Wissensgewinnungsprozess, wie bereits im Abschnitt 3.3 erwähnt, wird nach dem Vorgehensmodell des MESC von ITPL [SCHE17] durchgeführt.

Das folgende Kapitel beschreibt die KDD-Analyse auf produktionslogistische Firmendaten und die Erarbeitung von Optimierungspotenzialen durch den Einsatz von Data-Mining-Verfahren. Es folgt eine Zielbeschreibung mit anschließender allgemeiner Darstellung der Produktionssysteme der Elektronikfertigung. Nach einer Aufgabendefinition erfolgt die Datenauswahl mit der im Abschnitt 3.4 ausführlich erklärten Datenvorverarbeitung. Auf die strukturierten Daten werden verschiedene Methoden angewendet und miteinander verglichen, um am Ende eine passende Auswahl zu treffen, die Erkenntnisse am besten darstellen.

4.1 Zielbeschreibung

Das Ziel dieser Arbeit besteht erstens darin, dass allgemeine Potenzial der Wissensentdeckung im Produktionsumfeld der Elektronikfertigung aufzudecken. Bislang ist der Einsatz von Data-Mining in der Produktion sehr gering bzw. wird gar nicht in Anspruch genommen. Die Grundlage für die Anwendung ist aufgrund der anfallenden Datenmengen gegeben, jedoch wird sie nicht abgerufen und liegt ungenutzt in den Datenbanken. Trotz der Einschränkung des Schwerpunktes auf die Elektronikfertigung in dieser Arbeit, ist die Anwendung der Analyse auf allen Ebenen des Produktionsbereiches möglich.

Als zweites Ziel wird der Zusammenhang zwischen den Prozessparametern und Bauteilqualität mit Data-Mining-Verfahren untersucht, um somit Optimierungsmaßnahmen zu identifizieren. In der Elektronikfertigung werden elektronische Leiterplatten hergestellt, die bei WILO SE ein Bestandteil eines Pumpenmodules sind. Beim Reflow-Ofen handelt es sich um den Kernprozess der SMD-Fertigungslinie und bestimmt maßgeblich deren Qualität. Das Lötverfahren bestimmt die nachgelagerten Prozesse und somit die Nacharbeit und Ausschussmenge der Leiterplatten. Mit Hilfe von RapidMiner werden die Datenbestände untersucht, um Aussagen bezüglich der Wirkzusammenhänge zwischen Prozessparameter und Bauteilqualität zu treffen.

4.2 Produktionssystemaufbau der Elektronikfertigung

In diesem Abschnitt wird die Produktion von elektronischen Baugruppen mit ihren Fertigungsbereichen dargestellt und erläutert. Die vier Bereiche einer Elektronikfertigung gliedern sich in Warneingang/-ausgang, SMT-Fertigung, Nutzentrenner und THT-Fertigung.

- **Wareneingang/-ausgang**

  Alle Bestandteile, die für die Herstellung der Leiterplatten nötig sind, werden im Wareneingang geliefert und bereitgestellt. Hierbei werden die gelieferten Bauteile und unbearbeiteten Leiterplatten ausgepackt und überprüft. Bei der Qualitätskontrolle werden alle Bauteile einer Sichtkontrolle unterzogen, falls Auffälligkeiten zu sehen sind, wieder zurückgeschickt. Im Warenausgang werden die fertigen Leiterplatten dem Auftrag entsprechend im Vakuum verpackt und verschickt.

- **SMT- Fertigung**

  Der erste Arbeitsschritt für die unbearbeiteten Leiterplatten erfolgt im SMT-Fertigungsbereich (Surface Mounted Technology; dt.: Oberflächenmontagetechnik). In der SMT-Linie werden die Leiterplatten mit einem Datamatrix-Code (s. **Abbildung 21)** versehen, dieser wird anhand eines Lasers eingraviert. Dieser Code dient der Datenerfassung und für die eindeutige Rückverfolgbarkeit der Leiterplatte in der gesamten Wertschöpfungskette. Der nachfolgende Schritt in der SMT-Linie ist das Aufbringen der Lotpaste mit dem Schablonendruckverfahren. Anhand des Siebdruckverfahrens wird eine Schablone mit Aussparungen auf der Platte positioniert und Bohrungen bzw. Anschlussflächen mit der Paste verfüllt. Anschließend erfolgt mit dem SPI-System die Qualitätsprüfung wobei Kriterien wie Menge des Lots, Brückenbildung, Formfehler und Versatz überprüft werden.

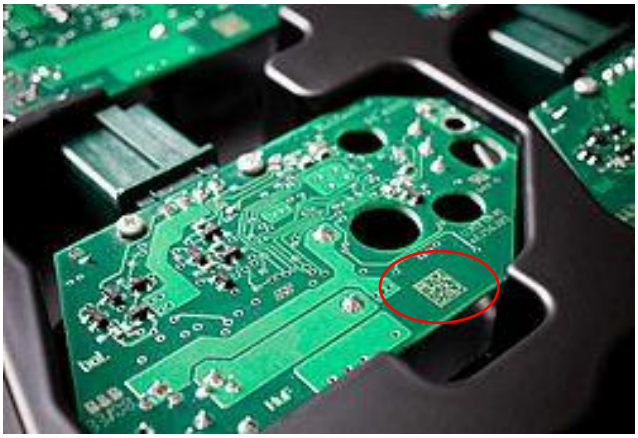


Abbildung 21: Leiterplatte mit eingraviertem Datamatrix-Code (Huf Electronics)

Die Bestückung der Leiterpatten erfolgt mit vollautomatischen Bestückungsautomaten, die mit Feedern (s. **Abbildung 22**) ausgestattet sind. Diese werden mit den benötigten Bauteilrollen ausgestattet und versorgen den Bestückungskopf (Picker) mit den entsprechenden Bauteilen. Die einzelnen Bauteile werden vom Picker angesogen und auf die Leiterplatte platziert.

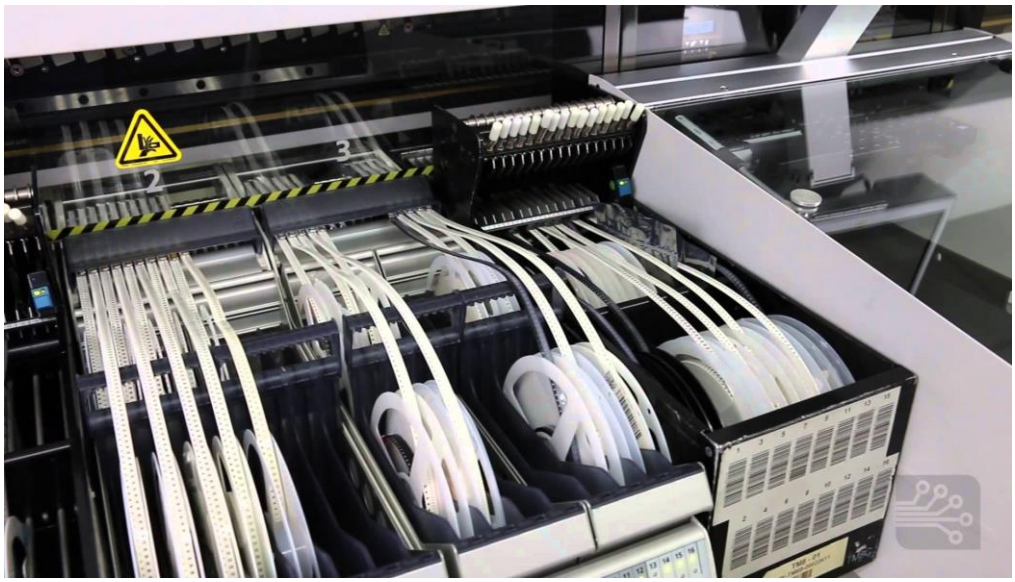


Abbildung 22: Bestückungsautomat (ESO Electronic)

Nachdem die Bauteile, ohne Drahtanschlüsse, auf den Platine platziert worden sind, erfolgt der Lötprozess im Reflow-Ofen (s. **Abbildung 23**). Bevor der Prozess startet, wird der Datamatrix-Code abgelesen und dies in der Datenbank für Prozess- und Qualitätsprüfungen hinterlegt. Um unnötig Speicherkapazität zu besetzen, wird der Code nur von der Hauptplatine (Nutzen) erfasst und nicht von jeder einzelnen Platine selbst. Die zuvor aufgetragene Lotpaste, welche aus einem Gemisch von Lotkugeln und Flussmittel besteht, hält aufgrund seiner klebrigen Eigenschaft die Bauteile fest und wird erst im Reflow-Ofen wieder leicht aufgeschmolzen. Die Leiterplatine durchlaufen sechs verschiedene Heizzonen, um das Lot auf die vorbestimmte Temperatur zu bringen und werden schließlich in der letzten Zone abgekühlt. Aufgrund der Erwärmung sammelt sich das flüssige Lot umlaufend an den Lötstellen an und vernetzt sich, wobei das flüssige Flussmittel verdampft. Die dabei wirkende Oberflächenspannung des geschmolzenen Lots, zieht die Bauteile auf die Mitte des Pads an. Die Selbstzentrierung wird durch die geringe Masse des Bauteils auf dem Pad begünstigt. Der komplette Lötvorgang erfolgt unter einer Stickstoff-Schutzgasatmosphäre und bestimmt maßgeblich die Benetzungseigenschaft des Lotes. Durch Einsatz von Stickstoff werden die nachteiligen

Eigenschaften des Sauerstoffs in der Anlage niedrig gehalten. Der Restsauerstoff in der Atmosphäre wird in *ppm* gemessen. Diese Atmosphäre begünstigt größere Benetzungskräfte und verhindern somit die Bildung von Metalloxiden, die zu Qualitätsverlusten führen.

Abbildung 23: Reflow-Ofen (Ersa)

Den letzten Schritt in der SMT-Linie stellt die AOI (Automatische optische Inspektion, s. **Abbildung 24**) dar. Dieses Prüfverfahren dient zur Kontrolle der Leiterplatten-produktion. Die Komplexität in der Platinenstruktur und die kleiner werdenden Bauteile erfordern stets eine zuverlässige Qualitätskontrolle, die ohne ein AOI nicht mehr zu bewältigen ist. Vor jeder Prüfung wird der Datamatrix-Code von jeder Platine ausgelesen und festgehalten, um die Rückverfolgbarkeit zu gewährleisten. Die Anzahl der Platinen auf den Nutzen kann variieren, dies hängt von der jeweiligen Größe ab. Um wirtschaftlich zu produzieren, besteht ein Nutzen aus mindestens sechs Platinen.

Die Prüfung der einzelnen Leiterplatinen erfolgt kamerabasiert und wird computergestützt mit einem hinterlegten Referenzbild abgeglichen. Anhand definierter Punkte werden die Platinen auf mechanisch defekte Bauteile, Bestückungsfehler, Einpressfehler und Qualität der Lötung überprüft. Falls die Platinen die Prüfung nicht bestehen, werden diese von einer Mitarbeiterin begutachtet und gegebenenfalls nachbearbeitet oder als Ausschuss sortiert.

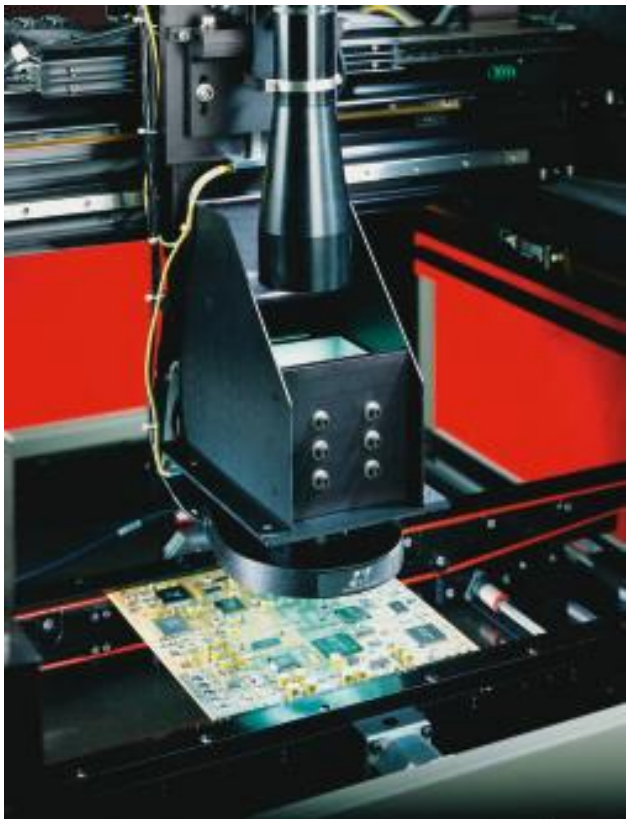


Abbildung 24: Automatische optische Inspektion (AOI) (EPP Industrie)

- **Nutzentrennzentrum**

  Die fertigen SMD-Platinen kommen zur weiteren Bearbeitung ins Nutztrennzentrum. Hierbei ist zu erwähnen, dass mit dem *Nutzen* das Zusammenfassen von mehreren kleineren Platinen auf eine große Platine bezeichnet wird. Im Trennzentrum werden die einzeln Platinen aus dem Nutzen, anhand verschiedener Trennverfahren wie Säge- und Fräseinrichtungen, getrennt (s. **Abbildung 25**). Die getrennten Leiterplatten werden zwischengelagert und nach dem First In-First Out (FIFO) Prinzip zur THT-Bestückung entnommen. Zur Vermeidung von Schmutz und Schadstoffeinflüssen sollten geöffnete Platinen möglichst schnell bearbeitet werden und in der Endmontage verbaut werden, deshalb gilt die Entnahme nur nach dem FIFO-prinzip.

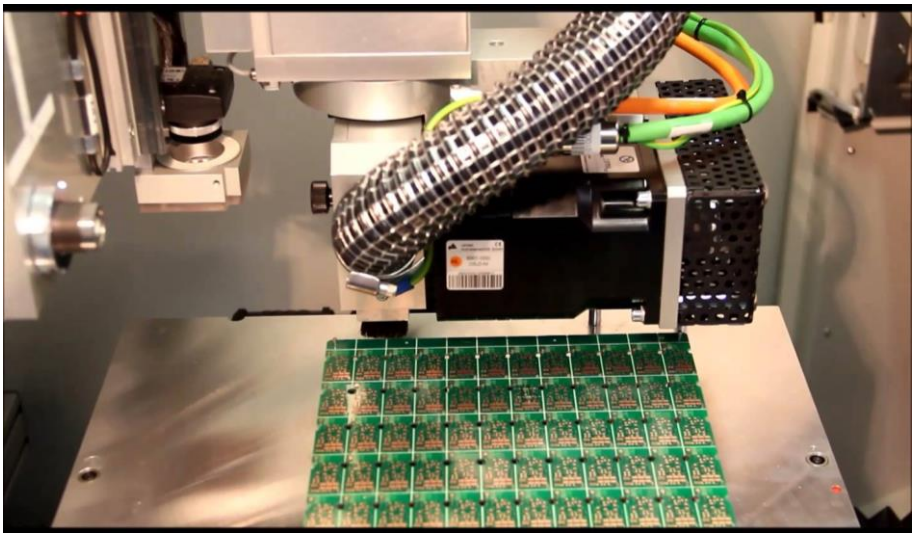


Abbildung 25: Nutzentrennzentrum (Systemtechnik Hölzer)

- **THT-Fertigung und Endmontage**

  In der THT-Fertigung werden Bauteile mit Drahtanschlüsse verbaut und stellen somit den letzten Schritt bei der Herstellung von Leiterplatten dar. Die Bauteile für die THT-Fertigung werden auf Rollen geliefert und in der Abteilung für Bauteilvorbereitung für die weitere Verwendung getrennt. Des Weiteren müssen die Anschlussdrähte für eine problemlose Bestückung geschnitten bzw. geformt und in Sichtlagerkästchen gefüllt werden. Die Bauteile werden je nach Ausstattung automatisch vom Picker oder von Mitarbeitern auf die Platine gesetzt. Davor ist zu beachten, dass die Platinen in einen Lötrahmen (s. **Abbildung 26**) gelegt werden, dieser sorgt für eine sichere und fehlerfreie Aufnahme in die Lötanlage.

Abbildung 26: Lötrahmen (LRT Technologie)

Nach der Bestückung werden die Leiterplatten im Lötrahmen auf einem Transportband zur Wellenlötanlage (s. **Abbildung 27**) transportiert. Die Platinen werden über ein oder zwei Lotwellen gefahren, wodurch ein flüssiges Lot gepumpt wird. Nach der Kühlung werden die Lötrahmen entfernt und die Platinen einem *In-Circuit-Test (ICT)* unterzogen. Dabei werden Fehler in der Leiterbahnführung, Lötfehler und Bauteil geprüft. Die fehlerfreien Produkte werden in der Endmontage verbaut oder für die Lieferung an die Kunden vorbereitet.

Abbildung 27: Wellenlötanlage (Ersa)

4.3 Aufgabendefinition und Datenauswahl

Die Datenanalyse in dieser Arbeit orientiert sich an dem Vorgehensmodell MESC, somit stellt dieser Abschnitt die ersten beiden Phasen des Modelles dar. Dabei findet eine Ausformulierung der Aufgabenstellung unter Beachtung von Rand- und Zielbedingungen sowie die Auswahl relevanter Daten statt.

4.3.1 Aufgabendefinition

Die erste Phase des MESC beschäftigt sich mit der konkreten Aufgabenstellung. Hier werden unternehmensspezifische Ziele, Erwartungen, Ressourcen und Restriktionen in einer Aufgabenstellung ausformuliert.

Aufgabenstellung (1.1)

Die Durchführung des KDD-Prozesses benötigt in der Praxis eine konkrete Aufgabenstellung. Die Frage dieser Arbeit bezieht sich auf den Zusammenhang zwischen Prozessparameter der Anlage und Bauteilqualität, die anhand verschiedener Methoden untersucht werden. Die Festlegung der Methode - bezüglich der Analyse von produktionslogistischen Daten - erfolgt am Ende durch den Vergleich von Ergebnissen.

4.3.2 Auswahl relevanter Datenbestände

Nach der Festlegung der Untersuchungsziele wird in der zweiten Phase des MESC der Zugang zu den relevanten Daten geschaffen und wichtige Datenbestände ausgewählt, die zum Erreichen des Ziels verhelfen.

Datenbeschaffung (2.1)

Die Daten für die Analyse kommen aus der Firma WILO SE und sind für die Arbeit zur weiteren Verarbeitung zur Verfügung gestellt worden. Die Datenbeschaffung erfolgte mithilfe der Experten im Unternehmen, somit kam von Anfang an eine Richtung definiert werden. Die Daten der Fertigung werden nicht komplett auf dem Server gespeichert und jedem zugänglich gemacht. Die verschiedenen Systeme wie das Manufacturing Execution (ME) oder AOI haben verschiedene Ansprechpartner und unterschiedliche Datenbanken, aus diesem Grund ist die Beschaffung der Daten schwierig und zeitintensiv.
Die Datentabellen stammen teilweise aus dem ME-System, beinhalten die Prozessparameter des Reflow-Ofens, und aus dem lokalen Server des AOI-System, die nach dem Lötvorgang die optische Qualitätsüberprüfung des Bauteils/-komponente durchführen. Aufgrund der Anbindung des Reflow-Ofens an das MES, sind die Daten zwar zentral einsehbar, aber nicht in eine Excel-Liste ohne Aufwand exportierbar. Der Datenexport erfolgte durch einen IT-Fachmann aus der

Zentrale, der nicht vor Ort anwesend ist. Hierbei ist zu erwähnen, dass das MES in der Einführungsphase ist und somit nicht alle Daten vollständig übertragen kann. Um ausreichend Datensatz für die Analyse zu bekommen, wird nach Absprache mit den Experten der Zeitraum auf September-Dezember 2017 festgelegt. Die AOI-Daten können nur vor Ort mithilfe des Mitarbeiters und mit erheblichen Zeitaufwand exportiert werden. Dies kann nur erfolgen, wenn die Produktionslinie während der Pause oder Wartung stillsteht.

Für die Analyse stehen mehrere Datenbestände zur Verfügung. Der erste Datenbestand „export_measurement_data“ enthält die Datentabelle der Prozessparameter des Reflow-Ofens und Lasermarkierung mit neun verschiedenen Attributen. Die Tabelle besteht aus 218413 Datenzeilen. Die Einteilung der Attribute erfolgt in „MEASURE_GROUP“, „MEASURE_NAME“, „ACTUAL“, „ORIGINAL_TEST_DATE_TIME“, „SFC“, „Material”, „Vorgang” und zwei weitere Spalten ohne Attributnamen. Die weiteren Datenbestände enthalten die AOI-Daten, die in mehrere Zeitabschnitte und Fertigungslinien unterteilt sind. Die beinhalten die folgenden Attribute wie „cModel”, „Fdate”, „BoardSN“, „T/B“, „Imulti“, „TestCount“ usw., die weiteren Attribute können in der Excel-Liste eingesehen werden.

Datenauswahl (2.2)

Die erste Aufgabe besteht darin, die nützlichen Daten für die Analyse auszuwählen, um den Data-Mining Prozess nicht mit einer zu hohen Datenmenge zu überlasten.

- **Auswahl der Datenbestände**
  Bereits zu Beginn wird die Möglichkeit in Anspruch genommen, sich mit den Experten im Unternehmen auszutauschen. Anhand dieser Besprechungen kann eine Einschränkung bezüglich des Analyseobjektes und der Daten erfolgen. Der Anschluss des Reflow-Ofens der SMD-Linie an das MES und die damit entsprechende Datenverfügbarkeit, ist die Analyse auf die SMD-Fertigung und auf den dazugehörigen Kernprozess eingeschränkt worden. Die Daten der Lasermarkierung werden in dieser Arbeit nicht weiter betrachtet. Der nachfolgende Prozess ist die Qualitätsprüfung der Bauteile anhand des AOI, für die Relation zwischen Prozess- und Qualitätsdaten ist dieser Prozess relevant und wichtig. Die Arbeit ist auf die Untersuchung der Prozessparameter des Reflow-Ofens und AOI-Daten beschränkt und die relevanten Daten sind dadurch ausgewählt und exportiert worden. Daher hat man die folgenden Datenbestände (s. **Tabelle 3**) erhalten. Hierbei ist zu erwähnen, dass mehrere Datenbestände des AOI-Systems existieren, aufgrund der systembedingten, automatischen Einteilung des Zeitraums und mehrerer Produktionslinien.

Tabelle 3: Genaue Beschreibung der Datenbestände

Prozess	Datenbestand
Reflow-Ofen	export_measurement_data
AOI	2175139_22.09.17-23.10.17 (Linie 2)
AOI	2175139_06.11.17-29.11.17 (Linie 2)
AOI	2175139_01.09.17-05.09.17 (Linie 3)
AOI	2175139_05.09.17-18.10.17 (Linie 3)
AOI	2175139_18.10.17-29.11.17 (Linie 3)
AOI	2175139_29.11.17-07.12.17 (Linie 3)

- **Auswahl der Datentabelle**

  Nach der konkreten Aufgabenstellung und Datenbeschaffung ist das Ziel, die Überführung der Daten in ein Zielformat. Die im letzten Abschnitt erwähnten Datenbestände besitzen jeweils eine Datentabelle mit den Prozessinformationen. Die vorhandenen Datentabellen stellen die Prozessparameter und die Qualitätsdaten für die Herstellung der Produkte dar, daher sind zunächst alle vorhandenen Tabellen relevant.

- **Auswahl der Daten**

  Die Tabelle „export_measurement_data“ beinhaltet die Prozessdaten von zwei verschiedenen Prozessschritten sowie von drei verschiedenen Produkten, die anhand der Materialnummer unterschieden werden. Die Verknüpfung der Anlagen an das MES ist noch nicht komplett abgeschlossen, insofern stehen nicht alle Daten zur Verfügung. In dieser Tabelle sind die Materialnummer 2175139 (Phoenix Para RK), 2175140 (Phoenix Para PWM) und die 2175141 (Phoenix Pico STG) vertreten. Hierbei handelt es sich um Leiterplatinen der Module für Wasserförderpumpen und gehören zu der neuen Generation im Produktportfolio. Für die Analyse werden mit den Prozessparametern zusätzlich auch die Qualitätsdaten benötigt, die jedoch nur für die Materialnummer 2175139 vorliegt.

Deswegen werden aus der Tabelle nur Informationen entnommen, die hauptsächlich die Materialnummer 2175139 und den Prozess SMD-Reflow-Ofen betreffen. Die AOI-Datenbestände beinhalten alle dieselben Attribute, nur sind diese in unterschiedliche Zeiträume eingeteilt, da das AOI-System die Datenbank nach Erreichung der Speicherkapazität die Datenbank erneuert.

In den folgenden Tabellen sind die wichtigen Attribute, der jeweiligen Datenbestände, mit den entsprechenden Beschreibungen aufgelistet. Die Auflistung erfolgt getrennt nach den Vorgängen, deshalb befinden sich in der **Tabelle 4** die wichtigen Attribute des Reflow-Ofens und in **Tabelle 5** vom AOI. Die Beschreibung dient für das Verständnis hinter dem Attribut und gibt die Bedeutung an. Allein mit dem Attributnamen vom Server ist kein eindeutiger Rückschluss auf den Vorgang bzw. Bedeutung ohne interne Produktionskenntnisse möglich. Die aufgelisteten Attribute spielen für die Verknüpfung der Tabellen als auch für die spätere DM-Analyse eine entscheidende Rolle.

Tabelle 4: Auflistung wichtiger Attribute des Reflow-Ofens

Attribut	Beschreibung
MEASURE_GROUP	Prozessvorgang
MEASURE_NAME	Prozessschritt
ACTUAL	gemessener Ist-Parameterwert
ORIGINAL_TEST_DATE_TIME	Zeitangabe
SFC	Produktseriennummer (PSN)
Material	Materialnummer

Tabelle 5:Auflistung wichtiger Attribute der AOI-Qualitätsdaten

Attribut	Beschreibung
cModel	Materialnummer
Fdate	Zeitangabe
BoardSN	Produktseriennummer (PSN)

Status	Qualitätsstatus
Errtype	Fehleranzeige

4.4 Datenvorverarbeitung

Um Data-Mining-Verfahren anwenden zu können, werden die unstrukturierten Daten einem Datenvorverarbeitungsprozess unterzogen. Um für die nachfolgende DM-Analyse vorzubereiten, werden verfahrensunabhängige Methoden des Datenvorverarbeitungsprozesses angewendet. Die Verarbeitung der Daten orientiert sich an das Vorgehensmodell von MESC und beinhaltet folgende Schritte: Formatstandardisierung, Gruppierung, Datenanreicherung und Transformation.

Formatstandardisierung (3.1)

Um das Data-Mining erfolgreich anwenden zu können, müssen alle relevanten Informationen bezüglich des Produktionsprozesses und Qualitätsprüfung in einer Haupttabelle enthalten sein. Das Ziel besteht darin, die Datentabellen in ein Standardformat zu überführen und die DM-Analyse zu ermöglichen.

- **Strukturänderung in der Prozessparametertabelle**
  In Bezug auf die spätere DM-Analyse weist die vorliegende Tabelle des Reflow-Ofens eine ungünstige Tabellenstruktur vor. In dieser Arbeit werden die Prozessparameter in Bezug auf die Qualität des Bauteils untersucht. Für eine effiziente Analyse ist es wichtig, dass die neunzehn verschiedenen Parameter unter dem Attribut MEASURE_NAME (s. **Tabelle 6**) jeweils ein separates Attribut bilden und horizontal in der ersten Zeile befinden. Die dazugehörigen Werte sind vertikal unter dem entsprechenden Attribut aufgelistet. Sobald die Fertigungsparameter ein separates Attribut bilden, wird das ursprüngliche Attribut MEASURE_NAME aufgelöst. Die restlichen Attribute bleiben unverändert und werden übernommen. Die Umstrukturierung erfolgt in diesem Fall anhand von Excel, weil die Änderung gut umzusetzen ist. In RapidMiner ist die Umsetzung der Struktur, wie in **Tabelle 7,** mithilfe von Operatoren aufwendig. Jedoch ist hier darauf hinzuweisen, dass in der Literatur - in Bezug auf Datenvorverarbeitung - eine Tabellenstrukturveränderung nicht explizit erwähnt wird, aber hier für nötig und wichtig gehalten wird, da die Analyse ansonsten nicht funktionieren wird.

Tabelle 6: Ausschnitt: Ursprungsstruktur der Datentabelle des Reflow-Ofens

MEASURE_GROUP	MEASURE_NAME	ACTUAL
SMT_OVEN_DC	Z1_TEMP_TOP	195
SMT_OVEN_DC	Z1_TEMP_BOT	129
SMT_OVEN_DC	Z2_TEMP_TOP	165
SMT_OVEN_DC	Z2_TEMP_BOT	145
SMT_OVEN_DC	Z3_TEMP_TOP	195
SMT_OVEN_DC	Z3_TEMP_BOT	168
SMT_OVEN_DC	Z4_TEMP_TOP	200
SMT_OVEN_DC	Z4_TEMP_BOT	177
SMT_OVEN_DC	Z5_TEMP_TOP	260
SMT_OVEN_DC	Z5_TEMP_BOT	260
SMT_OVEN_DC	Z6_TEMP_TOP	260
SMT_OVEN_DC	Z6_TEMP_BOT	260
SMT_OVEN_DC	Z7_TEMP_TOP	80
SMT_OVEN_DC	TRACK1_TRANSPORT_SPEED	900
SMT_OVEN_DC	TRACK1_TRANSPORT_WIDTH	240800
SMT_OVEN_DC	TRACK1_PCB_SUPPORT_WIDTH	120000
SMT_OVEN_DC	TRACK1_PCB_SUPPORT_HEIGHT	0
SMT_OVEN_DC	MACHINE_NITROGEN	1055
SMT_OVEN_DC	MACHINE_COOLING_BATTERY_TOP	1

Die gewünschte Struktur kann man nicht allein durch Transponieren der Zeilen erhalten, denn dabei bildet sich nur eine unendliche lange Zeile mit wiederholenden Attributen und hat nicht mehr den Charakter einer klassischen Tabelle. Die folgende Tabellenstruktur (s. **Tabelle 7**) ist durch den Einsatz von *Filtern*, *Kopieren und Einsetzen* in Excel ermöglicht worden. Die Filterung bezieht sich auf einen konkreten Wert oder Attribut und ermöglicht, die damit verbundenen Werte in der Spalte zu kopieren und in einer anderen Tabelle und Spalte wieder einzufügen. Dieser Vorgang ist bei allen Prozessparametern wiederholt worden. Nun ist jeder Parameter ein Attribut und die entsprechenden Werte befinden sich in derselben Spalte darunter.

Tabelle 7: Ausschnitt der Datentabelle nach Strukturänderung

Z1_TEMP_TOP	Z1_TEMP_BOT	Z2_TEMP_TOP	Z2_TEMP_BOT
195	129	165	145
195	129	165	145
195	129	165	145
195	129	165	145
195	129	165	145

Gruppierung (3.2)

In diesem Abschnitt werden die vorliegenden Datentabellen des Reflow-Ofens und AOI unter fachlichen Aspekten zusammengefasst und miteinander verknüpft. Die am Ende entstehende Haupttabelle wird für die DM-Analyse von Bedeutung sein.

- **Zusammenfassung der AOI-Datentabellen**
  In diesem Abschnitt geht es darum, verschiede Tabellen mit demselben Aufbau, Attributen und Information zu einer Haupttabelle bezüglich eines Systems zusammenzufassen. In dieser Arbeit gibt es bei den Qualitätsdaten, aufgrund unterschiedlicher Zeiträume, mehrere Tabellen. Die Reflow-Datentabelle besteht in diesem Fall aus einer Tabelle, wodurch die Zusammenfassung nicht nötig ist. Für die spätere Integration ist es wichtig, dass die Daten kompakt in einer Tabelle wiederzufinden sind. Dies kann klassisch mit Excel oder mit RapidMiner erfolgen. Mit Excel werden die Daten jeder einzelnen Tabelle, unter Beachtung der zeitlichen Angabe, ausgeschnitten und in einer neuen Tabelle hinzugefügt. Dieser Vorgang ist umständlich und erfolgt solange, bis am Ende nur eine Tabelle mit den AOI-Qualitätsdaten vorhanden ist. Mithilfe von RapidMiner geht die Zusammenfassung der Tabellen, anhand des Operators *Append*, schneller und müheloser.
  Die Prozessauslegung für die Zusammenfassung kann in **Abbildung 28** eingesehen werden. Zur Veranschaulichung sind an dem *Append-Operator* zwei Tabellen angeschlossen und der Ausgang mit einem *Write Excel-Operator* verbunden worden. Hierbei ist zu erkennen, dass der *Append-Operator* die Möglichkeit bietet, weitere Tabellen zu verknüpfen.

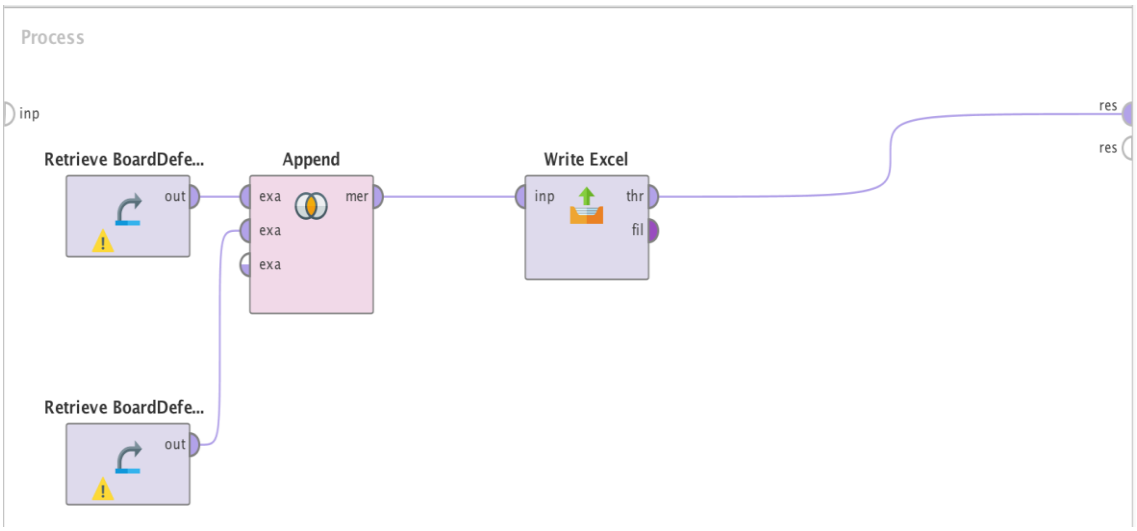


Abbildung 28: Prozessauslegung für die Zusammenfassung der AOI-Datentabellen

- **Datenintegration**

  Um Prozessparameter mit der Produktqualität untersuchen zu können, werden beide Datenbestände von unterschiedlichen Datenbanken in Relation gebracht. Durch sogenannte Merkmalsvektoren bzw. Marken wird der Zusammenhang zwischen den Beständen hergestellt (vgl. 3.4.1). Die markenbasierte Integration erfolgt anhand der PSN (Product Serial Number), die in allen Datentabellen vorhanden sind. Die PSN ist in der ursprünglichen Prozessparametertabelle unter dem Attributnamen „SFC“ und in der AOI-Tabelle unter „BoardSN“ zu finden. Das Problem für die Integration ist jedoch das Verfahren bei der Erfassung der Datamatrix-Code, wie bereits in Kapitel 4.2 erwähnt, wird im Reflow-Ofen nur der Code vom Nutzen erfasst und in der Datenbank hinterlegt. Das jede Platine selbst eine Datamatrix-Code hat, wird hierbei nicht beachtet. Dies hat zur Folge, dass in der Datentabelle, außer die PSN vom Nutzen, keine weiteren hinterlegt sind. Die AOI-Daten beziehen sich wiederum auf die einzelnen Platinen und erfassen bei der Qualitätsprüfung jede einzelne PSN mit und hinterlegen dies für die Rückverfolgbarkeit und weiterer Qualitätsbeurteilung im lokalen Server ab. Die **Tabelle 8** verdeutlicht die Darstellung der Materialgruppe und PSN in der Prozessparametertabelle, hierbei handelt es sich nur um die PSN des Nutzen.

Tabelle 8: Bsp. für die Erfassung der PSN vom Nutzen

Materialgruppe	PSN	Information bezüglich Datenzugehörigkeit
2175139	6003100708	PSN vom Nutzen
2175139	6003100715	PSN vom Nutzen
2175139	6003100722	PSN vom Nutzen
2175139	6003100729	PSN vom Nutzen

Um die Integration der beiden Hauptdatenbestände zu vollziehen, werden die jeweiligen PSN des Nutzens zusätzlich um sechs weitere PSN ergänzt. Dies wird anhand der **Tabelle 9** dargestellt. Die Umsetzung erfolgte mit dem Programm Excel, unter jeder Nutzen-PSN sind automatisch sechs Zeilen eingefügt und mit den zusätzlichen PSN ergänzt worden (grün hinterlegt). Für die Umsetzung werden Makros benötigt, manuell ist dies nicht machbar. Die für die Ergänzung der PSN eingesetzten Makros können im **Anhang 2** eingesehen werden. Die PSN der einzelnen Platinen ist eine aufbauende Zahlenreihe von der Nutzen-PSN um sechs Einheiten erweitert. Da es sich hier um eine große Datenmenge handelt, kann aus Zeitgründen nicht alles manuell eingefügt werden, daher werden hierzu erweiterte Excel-Kenntnisse benötigt, womit diese und die nachfolgenden Vorgänge der Tabellenverarbeitung drastisch vereinfacht werden können. Hierfür werden Makros geschrieben oder Funktionen formuliert, die trotz der

Datenmenge effizient arbeiten. Das Einfügen und Ausfüllen der zusätzlichen Zeilen kann mit RapidMiner nicht vollzogen werden, da für diese Vorgänge keine Operatoren existieren.

Tabelle 9: Bsp. für die ergänzte PSN der einzelnen Leiterplatinen

Materialgruppe	PSN	Information bezüglich Datenzugehörigkeit
2175139	6003100708	PSN vom Nutzen
2175139	6003100709	PSN von der einzelnen Platine
2175139	6003100710	PSN von der einzelnen Platine
2175139	6003100711	PSN von der einzelnen Platine
2175139	6003100712	PSN von der einzelnen Platine
2175139	6003100713	PSN von der einzelnen Platine
2175139	6003100714	PSN von der einzelnen Platine
2175139	6003100715	PSN vom Nutzen
2175139	6003100722	PSN vom Nutzen
2175139	6003100729	PSN vom Nutzen

Sobald das Hauptmerkmal des Datenbestandes ergänzt worden ist, ist die Verknüpfung beider Bestände möglich. Die PSN ist nun ein Primärschlüssel, womit Prozessdaten des Reflow-Ofens und die Qualitätsdaten vom AOI verknüpft werden können. Die Zuordnung kann mithilfe von Excel oder RapidMiner erfolgen. Hier bietet Excel Funktionen an, womit der Datenabgleich anhand des Primärschlüssel, hier die PSN, erfolgt und die benötigten Attribute mit seinen Zeilen in der neuen Tabelle ausgeben kann. Die Ausgabe der Zeile des Attributes erfolgt exakt in der Zeile, wo der Primärschlüsselwert übereinstimmt. Falls keine Übereinstimmung stattfindet, wird diese Zeile nicht ausgegeben und bleiben leer. Der Datenabgleich kann mit dem SVerweis oder IndexVergleich durchgeführt werden, jedoch ist zu beachten, dass die Funktionen aufgrund von Leerzeichen nicht direkt funktionieren, dafür werden die Werte gegebenenfalls geglättet. In Excel erfolgt die Verknüpfung mithilfe von Funktionen. In RapidMiner kann die Verknüpfung der beiden Tabellen mithilfe von Operatoren erfolgen. Den Ausgang für die Verknüpfung bilden die beiden Datentabellen, die anhand des *Retrieve-Operators* die Daten vom Repository aufrufen. Um die Verknüpfung mit dem *Join-Operator* zu vollziehen, werden für beide Datentabellen Schlüsselattribute festgelegt. Die Festlegung des Attributs zum ID-Attribut erfolgt mithilfe des *Set-Role-Operators*. In beiden Datentabellen ist die PSN ein entscheidendes Attribut, womit der Abgleich erfolgen und die Qualitätsdaten den Reflow-Daten zugeordnet werden kann.

Hier ist zu beachten, dass beide Schlüsselattribute denselben Attributennamen für den *Join-Operatoren* haben müssen, falls nicht erfolgt eine Fehlermeldung. Dies kann mit dem *Rename-Operator* behoben werden. An beiden *Retrieve-Operatoren* werden jeweils ein *Set-Role-Operator* verlinkt und die PSN als ID bestimmt. Die Verknüpfung beider Tabellen erfolgt mit dem *Join-Operator* und kann beliebig wieder in Excel ausgegeben werden. Die Prozessauslegung mithilfe der Operatoren kann in **Abbildung 29** eingesehen werden.

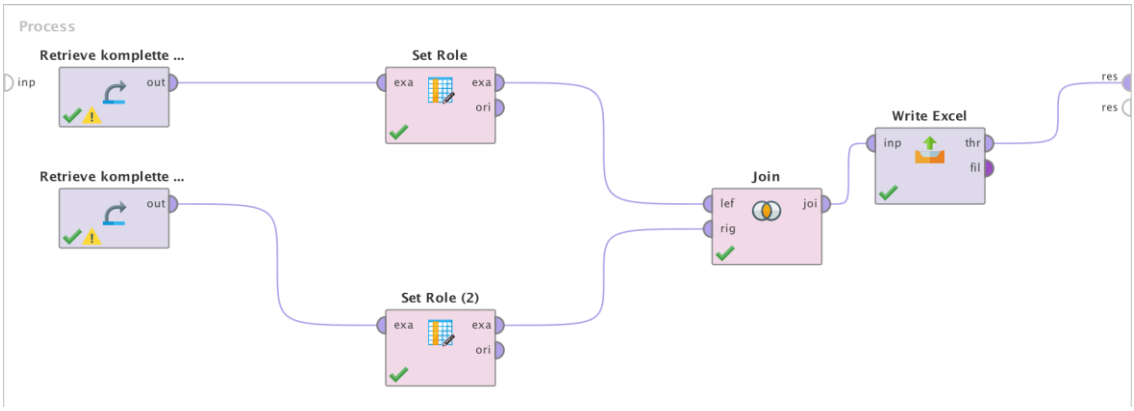


Abbildung 29: Verknüpfung der Tabellen mit dem Join-Operator

Datenanreicherung (3.3)

Die Datenanreicherung hat das Ziel den Datenbestand durch externe Informationen zu erweitern. Es werden neue Attribute generiert, um die Informationen effizienter zu erfassen. Die Analysedaten dieser Masterarbeit stammen direkt aus der Fertigung, hier werden keine zusätzlichen Daten von außen benötigt. Hierbei ist zu erwähnen, dass eine Datenanreicherung hinsichtlich der Generierung von neuen Attributen und Ergänzung von der PSN der einzelnen Platinen erfolgt ist. Bei den ergänzten Daten handelt es sich nicht um externe Informationen, die der Tabelle hinzugefügt worden sind. Die Generierung der neuen Attribute erfolgt nur mit den vorhandenen Informationen und anhand einer Änderung der Tabellenstruktur, in dem das ursprüngliche Attribut MEASURE_NAME aufgelöst worden ist. Die PSN der einzelnen Platinen ist eine logische und eine fest bedachte Folgerung von der PSN des Nutzens. Die Verwendung von bereits vorhandenen Daten, die Generierung von neuen Attribute und Ergänzung der PSN sind keine Informationen, die von außen kommen, deshalb nicht unter dem Abschnitt Datenanreicherung zu verfassen.

In der Fertigung wird gerne auf Benchmark-Analysewerte zurückgegriffen, um zu vergleichen, wie der Produktionsstand im Vergleich zum Wettbewerb ist. Daher ist die Produktion im Punkt Datenanreicherung nicht so stark vertreten und interessiert, wie das z.B. im Vertrieb der Fall ist.

Dort gehören Marktanalysedaten zu einem wichtigen Bestandteil, um die Prozesse, Strategien und Service weiter zu entwickeln.

Transformation (3.4)

Nach der Zusammenführung der Tabellen auf ein Zielformat, befinden sich nun alle Daten in einer Tabelle. Die Verknüpfung der Tabellen, die anhand des Primärschlüssels erfolgt, bringt mehr Datenattribute und -zeile mit sich. Um eine fehlerfrei und effiziente DM-Analyse durchführen zu können, müssen unnötige Attribute und redundante Daten entfernt werden. Im Abschnitt *Transformation* stehen Verfahren wie Aggregation, Dimensionsreduktion oder Stichprobenziehung sowie weitere Methoden zur Verfügung um eine Reduktion der Daten durchzuführen.

- ***Datenreduktion***

  In einer verknüpften Datentabelle findet man eine erhöhte Anzahl an Attributen und Datensätze. Dabei wird nicht jedes Attribut für die DM-Analyse von Bedeutung sein bzw. über nützliche Informationen verfügen. Die überflüssigen Attribute erschweren und verlangsamen den Data-Mining Prozess nur noch zusätzlich. Hierbei ergeben sich bei der Betrachtung der vorliegenden Tabellen Attribute, die weder in der Analyse noch für andere Auswertungen eine besondere Rolle spielen. Die häufigsten Gründe stellen die Attribute dar, die nicht gefüllt sind (leer, „0“, „NULL“, „???“) bzw. lediglich eine Ausprägung vorweisen. Des Weiteren sind Redundanzen unerwünscht, da sie lediglich keinen Informationsverlust darstellen, wenn man sie in der Datentabelle entfernt. Die Redundanzen tauchen in den vorliegenden Tabellen mit denselben Ausprägungen auf z.B. bei Zeitangaben, Materialangaben und Prozessvorgängen. Falls in den wichtigen Attributen, Ausreißer und fehlende Werte bekannt werden, werden diese Fehler behoben. In den folgenden Tabellen (s. **Tabelle 10** und **Tabelle 11**) werden die auffälligen Attribute der vorliegenden Datenbestände aufgelistet.

Tabelle 10: Auffällige Attribute in der Reflow-Ofen Datentabelle

Problemfall	Beschreibung	Namen des Attributen	Ausprägung
leere Attribute	ohne Attributenbezeichnung und nur mit einer Ausprägung		immer "A"
Ausprägung	Attribute nur mit einer Ausprägung	Material	„2175139“

verschiedene Attribute mit selber Ausprägung	selbe Ausprägung	MEASURE_GROUP Vorgang	„SMT_OVEN_DC“

Tabelle 11: Auffällige Attribute in den AOI-Qualitätsdaten

Problemfall	Beschreibung	Namen des Attributen	Ausprägung
Ausprägung	Attribute nur mit einer Ausprägung	T/B Test Count cModel	immer "T" immer "1" „2175139(2188502)_P06“
leere Ausprägung	keine Ausprägung vorhanden somit unbrauchbar	ErrCause	
Ausprägung	beschränkte Ausprägung	Imulti	1,2,3,4,5,6
Ausprägung	Attribut wird von einer Ausprägung dominiert	OFFx OFFy Distance The Data	größtenteils nur „0“

Die auffälligen Attribute, die in den beiden Tabellen ausführlich dargestellt worden sind, spielen in Bezug auf die Fragestellung keine besondere Rolle, da diese Attribute die Bauteilqualität gar nicht beeinflussen. Die Attribute stellen hier keine Prozessparameter dar, es sind lediglich zusätzliche Informationen wie Materialnummer, Reihenfolge der AOI- Aufnahmen oder Koordinaten der AOI-Kamera. Für die DM-Analyse können diese Attribute aus der Tabelle manuell entfernt oder anhand des Data-Mining Programm mit *Select Attributes* bzw. *Remove Useless Attribute*s aussortiert werden. Mithilfe von *Select Attributes* können die Attributen manuell vom Nutzer aussortiert werden, er entscheidet welche für seine Analyse von Bedeutung sind. Das *Remove Useless Attribut* entfernt automatisch nominelle Attribute, bei denen der häufigste Wert in mehr als dem angegebenen Verhältnis steht. Das Verhältnis wird durch den *nominal useless above* Parameter geregelt. Diese Eigenschaft kann bei Attributen zunutze gemacht werden, bei denen ein Wert alle anderen Werte dominiert. Falls nur ein Wert bzw. Werte mit kleinen Schwankungen in der Spalte auftauchen, werden diese Attribute entfernt, damit die DM-Analyse nicht beeinflusst wird.

Hierbei erfüllen beide Operatoren dieselbe Aufgabe, in dem die selektierten Attribute in der generierten Excel-Tabelle bzw. im Bereich „Results“ nicht mehr auftauchen und somit in der Analyse nicht weiter betrachtet werden. In **Abbildung 30** wird die

Verknüpfung des Retrieve-Operators an den *Select Attributes- bzw. Remove Useless Attributes-Operatoren* dargestellt. Beide sind unterschiedliche selbstständige Operatoren, deswegen werden beide Prozessstrukturen dargestellt. Bei Bedarf können beide Operatoren nacheinander verlinkt werden, falls z.B. nach dem *Remove Useless Attributes-Operatoren* noch einige Attribute vorhanden sind, die nicht weiter benötigt werden, werden diese anschließend mit dem *Select Attributes-Operatoren* entfernt (s. **Abbildung 31**). Der *Write Excel-Operator* ist zusätzlich eingefügt worden, um zu zeigen, dass der Output, falls die *Results*-Anzeige nicht ausreicht, in einer Excel-Liste wiedergegeben wird.

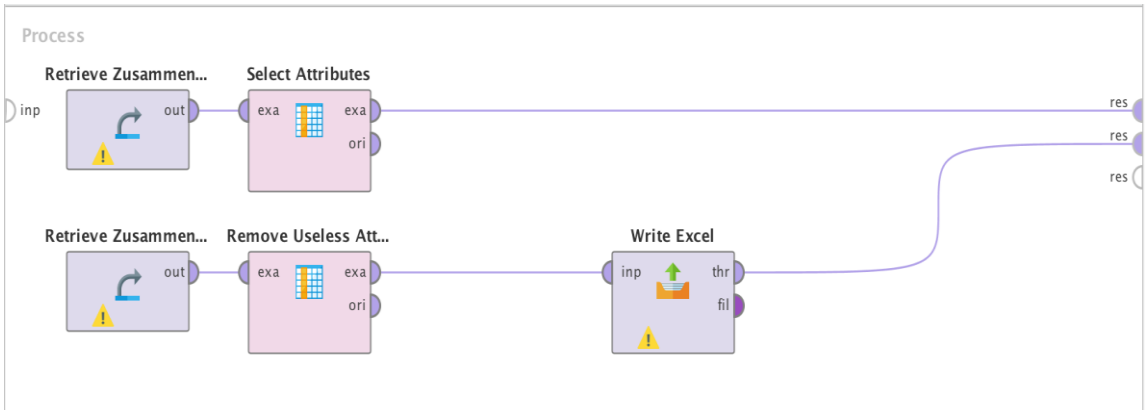


Abbildung 30: Operatoren für die Selektion von Attributen

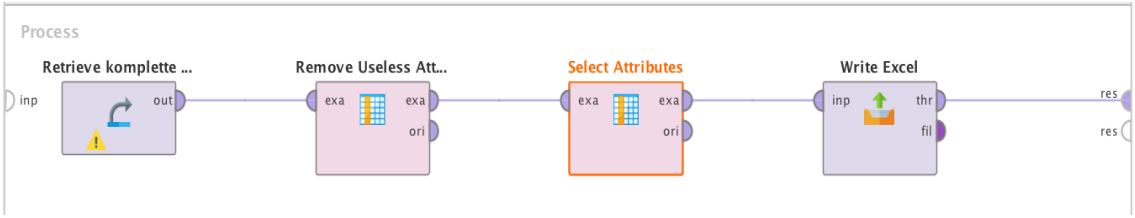


Abbildung 31: Zusammenschaltung der beiden Selektion-Operatoren

- **Änderung der Attributnamen**

  Nach erfolgreicher Selektion der nötigen Attribute durch die entsprechenden Operatoren erfolgt hierbei anschließend eine Änderung der Attributamen, um die Tabelle für alle verständlich zu machen. Unter den jeweiligen Attributnamen, die seit dem Datenexport aus den jeweiligen Anlagen und ME-System existieren, kann ohne Expertenwissen nicht wirklich nachvollzogen werden, was sich darunter verbirgt. Im Abschnitt 4.3.2 sind für die Analyse bereits, die für wichtig gehaltenen Attribute aufgegriffen und mit Fachbegriffen belegt worden. In der **Abbildung 32** wird die Prozessauslegung für die Änderung der Attributnamen mithilfe des *Rename-Operators* dargestellt.

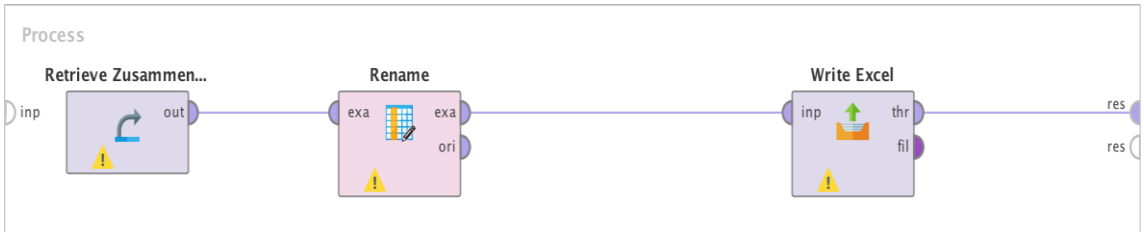


Abbildung 32: Prozessauslegung für die Änderung der Attributnamen

In der folgenden **Tabelle 12** werden die Attribute, nachdem eine Selektion erfolgt ist, mit neuen Attributnamen aufgelistet. Bei den aufgelisteten Attributen handelt es sich um Attribute, die für die Data-Mining-Analyse von Bedeutung sind. Hierbei wird kurz zum Verständnis in der Tabelle auf das ursprüngliche Attribut MEASURE_NAME eingegangen. Dieses Attribut beinhaltete beim Datenexport alle wichtigen Prozessparameter des Reflow-Ofens. Die Datentabelle wird hauptsächlich horizontal gelesen, hierdurch ist ein Zuordnen untereinander leichter möglich. Die vertikal aufgelisteten Prozessparameter sind aufgelöst und als neue Attribute integriert worden. Aufgrund dieser Änderung wird auf der rechten Tabellenhälfte der Hinweis gegeben, dass es sich um ein neu generiertes Attribut handelt.

Tabelle 12: Vollständige Attribute der Endtabelle mit den geänderten Attributnamen

ursprünglicher Attributname	neuer Attributname
Material, cModel	Materialnummer
SFC, BoardSN	PSN
MEASURE_NAME	wurde aufgelöst
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	Z1_Temp_Top
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	Z1_Temp_Bot
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	Z2_Temp_Top
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	Z2_Temp_Bot
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	Z3_Temp_Top
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	Z3_Temp_Bot

Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	Z4_Temp_Top
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	Z4_Temp_Bot
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	Z5_Temp_Top
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	Z5_Temp_Bot
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	Z6_Temp_Top
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	Z6_Temp_Bot
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	Z7_Temp_Top
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	MACHINE_NITROGEN
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	TRACK1_TRANSPORT_SPEED
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	TRACK1_PCB_SUPPORT_WIDTH
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	TRACK1_PCB_Support_HEIGHT
Attribute neu erstellt, aufgrund der Auflösung von „MEASURE_NAME	MACHINE_COOLING_BATTERY_TOP
Status	Qualitätsstatus

4.5 Durchführung des Data-Mining auf produktionslogistische Daten

In diesem Abschnitt wird ein geeignetes Data-Mining-Verfahren für die produktionslogistischen Daten gesucht und angewendet um entsprechend die Fragestellung beantworten zu können. Dabei stehen nach dem MESC-Vorgehensmodell noch einige Phasen offen, die in diesem Kapitel bearbeitet werden. Nach der Festlegung der Aufgabendefinition, Auswahl der Daten und Datenvorverarbeitung besteht die Vorbereitung und Anwendung des Data-Mining-Verfahren sowie eine anschließende Bewertung der Ergebnisse an. Die Auswertung am Ende verleitet zum weiteren Kapitel, in dem dort Handlungsempfehlung vorgeschlagen werden.

4.5.1 Vorbereitung des Data-Mining-Verfahren

Zur Durchführung des DM-Verfahrens wird in diesem Abschnitt die Methode und das DM-Werkzeug ausgewählt, die eine entscheidende Rolle in der Wissensgewinnung einnehmen. Dafür wird das Data-Mining-Programm nochmal aufgegriffen und Besonderheiten vorgestellt. Des

Weiteren besteht die Aufgabe, die passende Methode für produktionslogistische Daten auszuwählen, um diese im Abschnitt 4.5.2 anzuwenden.

Verfahrensauswahl (4.1)

Um nach der Datenvorverarbeitung mit den vorliegenden Daten Wissen zu generieren, sind Data-Mining-Verfahren essentiell. Sie beeinflussen und prägen am Ende das Wissensmuster und die daraus schließende Interpretation und Wissensgewinn. Bei der Verfahrensauswahl wird generell eine Vorentscheidung bezüglich der Methode getroffen, womit die Daten analysiert werden.

Im Abschnitt 3.5.2 wurden die in Frage kommenden Methoden für produktionslogistische Daten ausführlich vorgestellt. In wieweit sich die Methoden für die Fragestellung bewähren ist nicht absehbar und deshalb wird in diesem Abschnitt noch nicht auf eine eindeutige DM-Methode festgelegt. Die drei vorgestellten Methoden, in diesem Fall Entscheidungsbaum, Clustern und Assoziationsanalyse, werden unabhängig voneinander auf die vorverarbeiteten Daten angewendet und Ergebnisse miteinander verglichen um somit eine Methodenauswahl zu treffen. Dabei ist entscheidend, welches Verfahren die plausibelsten Ergebnisse liefert.

Werkzeugauswahl (4.2)

Für die Umsetzung von DM-Analysen existieren, wie bereits in dem Abschnitt 3.5.3 vorgestellt, viele Standardsoftware. RapidMiner gehört zum Teil dieser Standardsoftwares hinzu und wird aufgrund seiner einfachen Bedienung und Prozessauslegungsmöglichkeiten gerne bevorzugt benutzt, weil damit sogar unnötige Programmierungen vermeidet bzw. Programmierkenntnisse nicht benötigt werden. Es stehen zahlreiche Operatoren und hilfreiche Informationen zur Verfügung um beliebige Prozessschritte zu erstellen.

Fachliche und technische Kodierung (4.3-4.4)

Die technische bzw. fachliche Kodierung wird in diesem Abschnitt sich eher nur auf den Datentyp beziehen. Jedes Verfahren arbeitet mit anderen Datentypen zusammen, dementsprechend müssen sie an das Verfahren angepasst werden. Deshalb ist es wichtig zu wissen, dass die drei Verfahren unterschiedliche Bedingungen an den Datentyp stellen und diese für einen reibungslosen Ablauf sichergestellt werden müssen.

Die Datensätze wurden beim Importieren auf den Datentyp polynominal eingestellt und im RapidMiner gesichert. Der *Entscheidungsbaum mit dem ID3-Algorithmus* arbeitet mit dem Datentyp ohne Probleme, daher sind keine Veränderungen vorzunehmen. Auch generell sind beim ID3 keine festen Datentypen vorgegeben bzw. fest definiert. Die *Clusteranalyse mit dem k-Means Algorithmus* dagegen arbeitet nur mit numerischen Daten, daher ist es nötig den Datentyp von polynominal auf numerisch zu ändern. Dies kann durch den Operator *Nominal to Numerical* vollzogen werden. Die Assoziationsanalyse mit dem FP-Growth- Analyse arbeitet nur mit dem

Datentyp binominal zusammen. Daher ist einer Änderung von polynominal nach binominal erforderlich. Den Datentyp kann man anhand des Operators *Nominal to Binominal* umändern und den Algorithmus starten. Diese Maßnahmen sind für die optimale Anwendung der Methoden wichtig und sollte beachtet werden.

4.5.2 Anwendung des Data-Mining-Verfahren und Weiterverarbeitung der Ergebnisse

In diesem Abschnitt werden die zuvor festgelegten Methoden auf die vorverarbeiteten Daten angewendet. Damit dies erfolgen kann werden die Verfahren anhand von RapidMiner modelliert. Die genaue Festlegung auf eine Methode hätte bereits in Abschnitt 4.5.1 unter Verfahrensauswahl (4.1) erfolgen sollen, aber die Entscheidung wird erst nach der Modellierung und Ausführung getroffen. Die Methoden, die hier angewendet werden, sind in Abschnitt 3.5 ausführlich beschrieben. Für die Anwendung im Data-Mining wurde auf drei bekannte und häufig eingesetzte Methoden festgelegt, hierbei handelt sich um den Entscheidungsbaum, Clustern und Assoziationsanalyse.

Hierbei werden nach dem Vorgehensmodell MESC folgende Schritte abgearbeitet: *Entwicklung eines Data-Mining-Modells (5.1) sowie das Training des entwickelten Modells (5.2)*. Es wird keine strikte Trennung zwischen den beiden Schritten ergeben, weil dies nicht für sinnvoll gehalten wird. Beide Schritte werden zusammen bearbeitet und dargestellt.

Modellierung und Training von Data-Mining-Modells (5.1 und 5.2)

- **Allgemeine Modellierung der Datenvorverarbeitung**

  Für die Modellierung stellt die Software RapidMiner verschiedene Operatoren wie zum Beispiel den *Join, Append* und zum Testen *Apply Model* und Performance zur Verfügung. Zur Entwicklung des Data-Mining-Modell ist es erforderlich, dass die zuvor verarbeiteten Daten in einer zusammen gefassten und verknüpften Enddatentabelle existieren. Deshalb dient dieser Abschnitt der Modellierung zur Vorbereitung der Daten für die nachfolgende Analyse und dient als wichtiger Prozessbaustein und ist an allen Methoden anknüpfbar (s. **Abbildung 33**). Die Datentabellen des Reflowofens und AOI-Systems werden anhand des *Retrive Operator*, der den importierten Datensatz enthält, bereitgestellt. Die Datentabellen vom gleichen Typ werden mit dem *Append-Operator* zusammengefasst und innerhalb des *Subprocess-Operator* verlagert. Der Subprocess-Operator führt einen Prozess innerhalb eines Prozesses aus, somit können Prozesse verlagert werden und im Hauptfenster die Operatoren übersichtlich strukturiert werden. Der Join-Operator

verbindet anhand des Schlüsselattributes beide Datentabelle zu einer Datentabelle mit dem entsprechenden Prozessparameter und Qualitätsstatus, die für die Analyse entscheidend sind. Damit der Join-Operator überhaupt seine Funktion vollbringen kann, ist es erforderlich, dass zwei entscheidende Attribute - PSN - denselben Attributnamen haben. Mithilfe des *Rename-Operators* erfolgte die Umbenennung von *BoardSN* zu *PSN*. Anschließend nach der Zusammenfassung erfolgt die Reduktion unnötiger Attribute, die für die bevorstehende Analyse keinen Mehrwert bieten und somit entfernt werden können. Die Entfernung unnötiger Attribute entlastet den DM-Prozess und macht die Analyse übersichtig. Bei den AOI-Attribute sind bis auf den *Qualitätsstatus* unnötige Attribute dabei, die eher für die technische Instandhaltung der AOI-Kamera von Bedeutung sein würde, als für die Beurteilung der Qualität des Bauteils. Der Operator *Remove Useless Attributes* macht bereits eine Vorselektierung, indem Attribute, die konstante Werte oder komplett leere Zeilen aufweisen und keinen Mehrwert für die Analyse bieten, entfernt werden. Die Parameter des *Remove Useless Attributes* wurden nicht verändert und bei der Standardeinstellung belassen. Die genaue Prozessauslegung des *Append* und *Join* Operators können im Abschnitt 4.4 eingesehen werden.

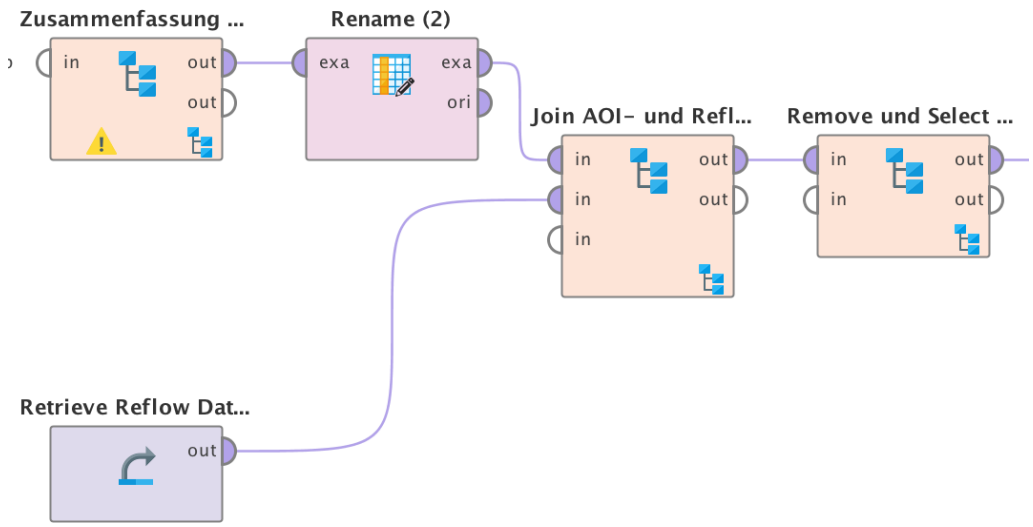


Abbildung 33: Modellierung der Datenvorverarbeitung

In **Tabelle 13** sind die Prozessparameter mit ihren jeweiligen Werten aufgelistet. Dabei wurden die Temperaturen der einzelnen Heizzonen aufgelistet, um zu zeigen wie weit die Spanne der Abweichung verlaufen und zu Verdeutlichung die Differenz gebildet. Bei dem Parameter *MACHINE NITROGEN*, gibt den Restsauerstoffgehalt an, waren die Messwerte unterschiedlich hoch, dass sie nicht einzeln aufgelistet werden konnten. Die vier anderen Parameter *TRACK SUPPORT WIDTH, TRACK1 SUPPORT HEIGHT,*

MACHINE COOLING BATTERY, *TRACK1 TRANSPORT SPEED* sind im gesamten Produktionsprozess konstant und spielen ebenfalls keine besondere Rolle in der Analyse. Die Abweichungen sind bei einigen gleich 0, somit ist dieser Wert konstant im gesamten Prozess. Daher kann man dieses Attribut selektieren bzw. entfernen, weil damit die Bauqualität nicht beeinflusst wird. Die achtzehn Parameter sind für eine Analyse, die recht übersichtlich sein soll, relativ hoch. Aus diesem Grund werden Parameter mit der Abweichung 0,1 und 2 selektiert, weil eine Abweichung von maximal 2 Grad als Toleranzgrenze betrachtet werden kann und keine besonderen Einfluss auf die Qualität des Produktes ausübt. Alle Abweichungen größer 2 Grad und der Restsauerstoffgehalt in der Atmosphäre (MACHINE NITROGEN) weisen drastische Abweichungen (rot markiert) auf, die nicht toleriert werden und als Inputfaktoren für die Analyse dienen um die Beziehung zwischen diesen Parameter und Bauteilqualität zu überprüfen. Die Selektierung von Attributen, die keinen großen Einfluss ausüben, führen zu kürzeren Laufzeiten und entlasten das System.

Tabelle 13: Prozessparameter und ihre Abweichungen

Prozessparameter							**Abweichungen Δ**
Z1 TOP	195	196					1
Z1 BOT	129	130	131	132	133	134	5
Z2 TOP	165						0
Z2 BOT	145	146	147				2
Z3 TOP	195	196					1
Z3 BOT	167	168	169				2
Z4 TOP	200	201					1
Z4 BOT	175	176	177	178	179		4
Z5 TOP	260						0
Z5 BOT	258	259	260	261			3
Z6 TOP	260						0
Z6 BOT	259	260	261				2
Z7 TOP	80	81					1
MACHINE NITROGEN	409	1294					hohe Abweichungen

TRACK SUPPORT WIDTH	12000						0
TRACK1 SUPPORT HEIGHT	0						0
MACHINE COOLING BATTERY	1						0
TRACK1 TRANSPORT SPEED	900						0

- **Entscheidungsbaum mit dem ID3-Algorithmus**

  Die erste Methode, deren Modellierung ausführlich dargestellt wird, ist der Entscheidungsbaum mit dem ID3-Algorithmus. Die zusammen gefassten Datensätze durch den Join-Operator spielen für die weitere Analyse eine besondere Rolle, denn diese dienen als Input für den ID3-Algorithmus. Um die aufgelisteten Attribute besser nachzuvollziehen, werden anhand des *Rename-Operators* verständliche Namen zugewiesen. Dieser Prozessschritt ist für den ID3 Algorithmus nicht entscheidend, aber hier nur eine formale Entscheidung. Damit der Algorithmus des Entscheidungsbaumes erst überhaupt funktionieren kann, ist die Zuweisung eines Label-Attributes, das Zielattribut, notwendig. Der Qualitätsstatus des Produktes nimmt eine entscheidende Funktion ein, deshalb wird dies als Label-Attribut, mithilfe des *Set Role*-Operators, festgelegt.

  Nach der Festlegung des Label-Attributen werden die Daten an den Validierungsoperator weitergegeben (s. **Abbildung 34**) Innerhalb des Validierungsblocks befinden sich zwei Subprozesse, auf der rechten Seite der Trainings- und auf linken Seite der Testblock (s. **Abbildung 35**) . Im Block erfolgt eine Aufteilung der Datensätze, weil für das Training und Testen separate Datensätze benötigt werden. Die nötigen Datensätze werden innerhalb des *Cross-Validation*-Operators, in mehreren Gruppendatensätzen unterteilt und zufällig ausgewählt. Die Anzahl der Gruppendatensätze wird vom Anwender im Operator Cross Validation unter „Paramters; number of folds“ individuell bestimmt. Bei klassischen train-and-test Methoden liegt der Anteil größtenteils bei 70 % Trainingsdaten- und 30 % Testdatensätze [ESTE00]. Die Trainingsdaten überwiegen die Testdaten. In der Kreuzvalidierung werden die Daten in die Gruppendatensätze, die vom Anwender bestimmt wurden, unterteilt und jede Gruppe wird einmal als Testdatensatz verwendet, die restlichen als Trainingsdaten eingesetzt. Somit ist gesichert, dass keine Daten weder fürs Training, als auch für das Testen des Modells verloren gehen. Innerhalb des Validierungsblocks werden der *ID3* und das *Apply Model* verbunden. Der ID3 ist der Klassifizierungsalgorithmus und nutzt die Trainingsdatensätze für die Erstellung des

Modells. Der ID3-Operator bietet verschieden Parameter an, die zur Optimierung des Entscheidungsbaumes beitragen und individuell eingestellt werden können, worauf nicht weiter eingegangen wird. Sowohl das Modell vom *ID3*-Operator als auch die Testdatensätze, werden dem *Apply Model*-Operator übergeben. Die Testdatensätze werden gegen das übergebene Modell getestet. Für die Überprüfung der Genauigkeit wird der *Performance*-Operator verwendet und wertet die Ergebnisse schließlich aus. Anschließend werden die Ergebnisse in *Results* präsentiert. Der komplette Prozess gibt als Ergebnis das Modell, die Datensätze und den Mittelwert der Genauigkeit (average accuracy) an.

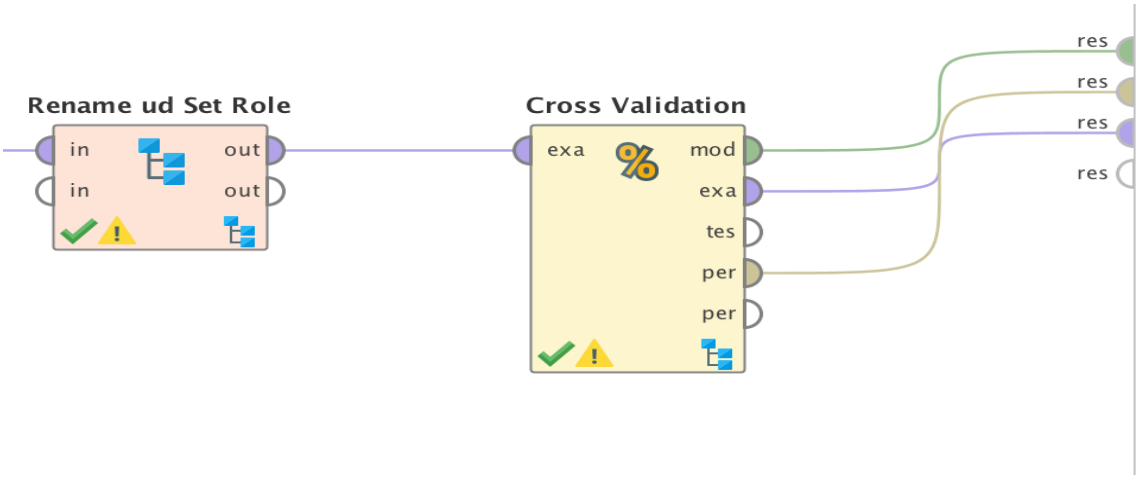


Abbildung 34: Modellierung des ID3-Algorithmus mit Cross-Validation

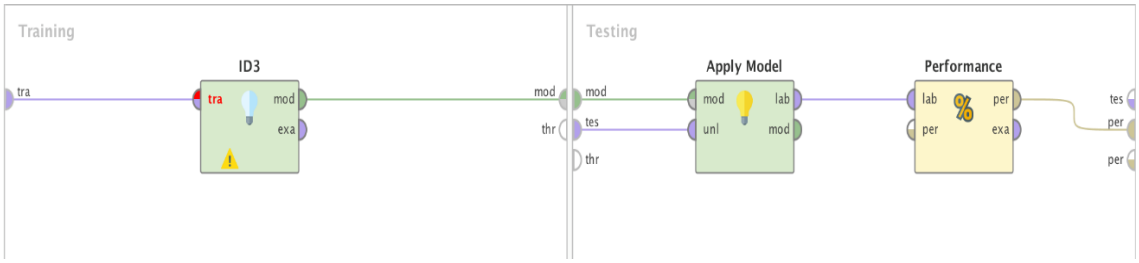


Abbildung 35: Innerhalb des Validierungsblocks

- **Clusteranalyse mit dem k-Means-Algorithmus**

In diesem Abschnitt geht es um die Modellierung des k-Means Algorithmus der Clusteranalyse. Die Festlegung auf den k-Means Algorithmus erfolgt aufgrund der geringen Iterationen für die Erzeugung von Clustern und der überschaubaren Darstellung. In **Abbildung 36** kann der wichtige Part der Clusteranalyse eingesehen werden, sie schließt sich der Datenvorbereitung an. Die Prozessauslegung erfolgt anhand drei wichtiger Operatoren. Die Grundvoraussetzung für den Einsatz des k-Means Algorithmus ist der numerische Datentyp. Je nach vorliegendem Datentyp kann der entsprechende Operator im RapidMiner ausgewählt werden um den Datentyp zu ändern. In diesem Fall haben die Datensätze den polynominalen Datentyp, daher kommt nur der Operator *Nominal to Numerical* in Frage und wird in den Prozess eingefügt. Dieser Operator wird zwischen den *Rename* und *Clustering* Operatoren eingefügt. Die Hauptanalyse wird erst durch den Operator k-Means ermöglicht. Um die Genauigkeit des Clustermodells zu überprüfen, wird zusätzlich der Operator *Cluster Distance Performance* an den k-Means Operator angeknüpft. Die Validierung des Prozesses gibt schließlich an, wie groß die Entfernung der Datensätze vom Centroiden liegen. Die Ergebnisse können in *Results* eingesehen werden.

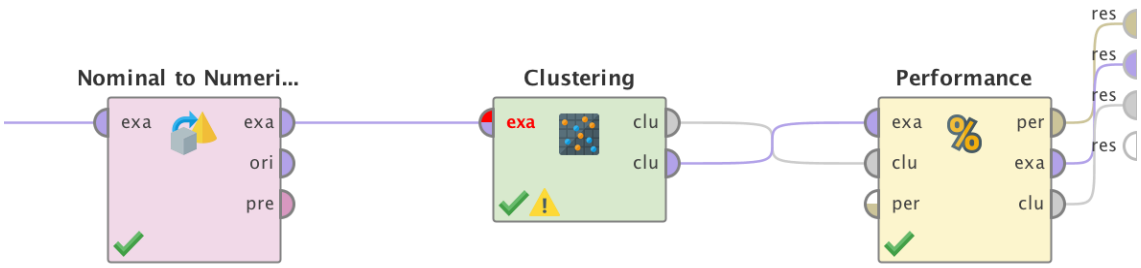


Abbildung 36: Clusteranalyse mit dem k-Means Algorithmus

- **Assoziationsanalyse mit FP-Growth-Algorithmus**

Zum Schluss wird die Modellierung des FP-Growth Algorithmus vorgestellt. Genau wie bei der Clusteranalyse arbeitet das Verfahren nur mit einem bestimmten Datentyp. Hierbei handelt es sich um den Datentyp *Binominal*. Je nach vorliegendem Datentyp kann auch hier der entsprechende Operator im RapidMiner ausgewählt werden, um entsprechend den Datentyp zu ändern. Die Datensätze weisen den polynominalen Datentyp, daher wird der Operator *Nominal to Binominal* gebraucht und deshalb in den

Prozess eingefügt. Für das Verfahren der Assoziationsanalyse wurde das FP-Growth Algorithmus ausgewählt, die in Abschnitt 3.5.2 ausführlich dargestellt wurde. Dieser wird zwischen den Operatoren *Nominal to Binominal* und *Create Association Rules* positioniert und angeschlossen. Die generierten frequent item sets werden von dem Operator *Create Association Rules* übernommen und entwickelt damit Assoziationsregeln, die am Ende als Ergebnis ausgegeben werden. Im FP-Growth Algorithmus können entscheidende Parameter wie minimale Anzahl an Itemsets sowie den minimalen Supportwert beliebig angepasst werden. In **Abbildung 37** wird die Prozessauslegung für den FP-Growth Algorithmus dargestellt.

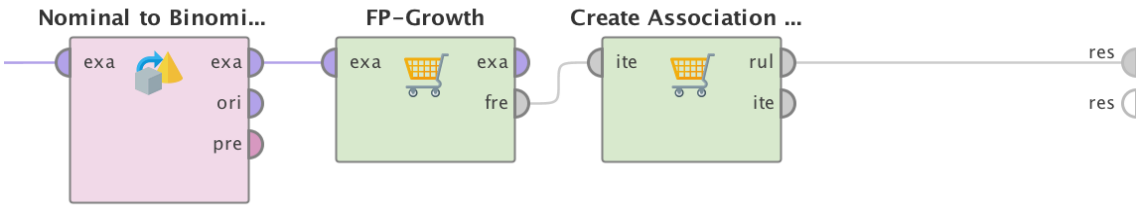


Abbildung 37: Assoziationsanalyse mit dem FP-Growth Algorithmus

4.6 Darstellung der Ergebnisse und Bewertung der Prozesse

Nach Modellierung und Anwendung der einzelnen Verfahren werden in diesem Abschnitt die Ergebnisse dargestellt und bewertet. Die Festlegung auf eine Methode wird nach Auswertung der Ergebnisse entschieden, welche für die Auswertung der Prozessparameter in Frage komme und plausible, anwendbare Ergebnisse anzeigen.

Darstellung und Bewertung der Ergebnisse vom ID3-Entscheidungsbaum

In diesem Abschnitt werden die Ergebnisse, die mit dem ID3-Entscheidungsbaum generiert wurden, dargestellt und bewertet.

Das besondere an einem Entscheidungsbaum ist, dass ein Zielattribut als Label-Attribut definiert und dies im Blattknoten wiedergibt. In dieser Arbeit werden Prozessparamater des Reflow-Ofens mit der Qualitätsausprägung des Produktes in Relation gesetzt. Es werden mögliche Kombinationsmöglichkeiten angezeigt, die zur gewünschten Ausprägung (Zielattribut) führen. Die Analyse von mehreren Parameterattributen belasten das System stark und verhindern teils den Abschluss der Analyse. Der Prozess benötigt eine höhere Speicherkapazität, die der Rechner nicht zu Verfügung stellen kann. Sobald die Anzahl der Attribute reduziert wird, verläuft die Analyse flüssiger und Ergebnisse können dargestellt werden.

Untersuchung von Entscheidungsbaumstrukturen anhand von Parametervariationen

Für die Datenauswertung mit Entscheidungsbäumen ist der Aufbau des Baumes sehr entscheidend. Deswegen wird in diesem Abschnitt der Aufbau näher betrachtet, indem Parameter variiert werden. Die Variation wird an folgenden Parameter durchgeführt: „minimal gain, criterion und number of folds“ In **Abbildung 38** wird ein Teilausschnitte des Entscheidungsbaums mit mehreren Attributen präsentiert. Die Parameter wurden folgend ausgewählt „criterion: information_gain; number of folds: 10; minimal gain: 0.1“.

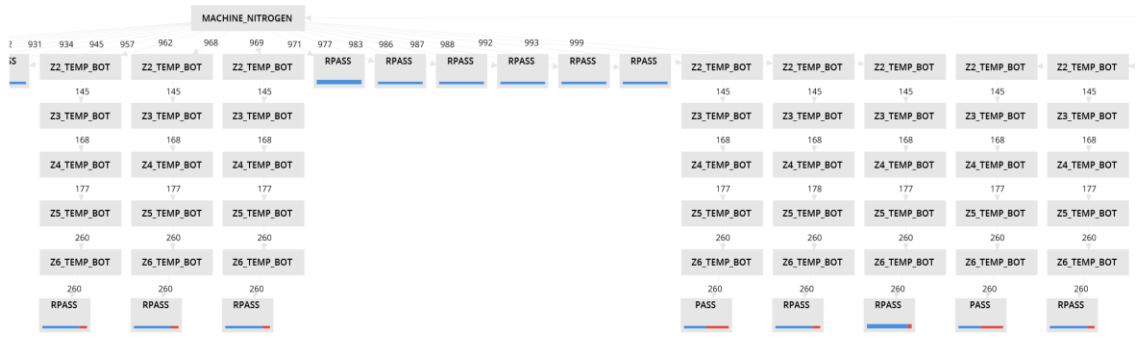


Abbildung 38: Teilausschnitt vom Entscheidungsbaum-Modell mit mehreren Attributen

Ein Parameter bildet den Wurzelknoten des Baumes und leitet die nächsten Kanten und inneren Knoten ein um das Zielattribut zu erreichen. Der Wurzelknoten, hier das Attribut Machine Nitrogen, bildet aufgrund der starken Variation viele Kanten, um die komplette Abweichung bzw. jeden einzelnen Wert darzustellen. Die Baumstruktur wird in die Breite gezogen und wird unüberschaubar groß. Auch die Selektierung von Attributen machen den Entscheidungsbaum nicht übersichtlicher und einfacher. Hier wurden Attribute mit der höchsten Abweichung ausselektiert. In **Abbildung 39** wird der Entscheidungsbaum dargestellt. Hierbei ist nur die Anzahl der inneren Knoten reduziert wurden.

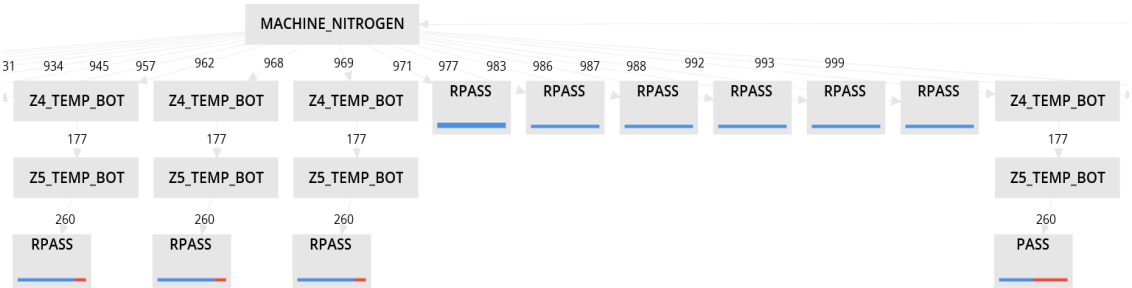


Abbildung 39: Teilausschnitt des Entscheidungsbaum-Modell mit den Attributen der höchsten Abweichungen

Bei Änderung des Parameters „criterion“ oder „minimal gain“ kann festgestellt werden, dass die Baumstruktur nicht mehr die ursprüngliche Form beinhaltet, wie bei den vorherigen Abbildungen. Die Attribute „Machine Nitrogen“ bildet in dem Entscheidungsbaum nicht mehr den Wurzelknoten und kommt erst in der mittleren Struktur vor (s. **Abbildung 40**). Es wird als zentrales Attribut bei dieser Analysemethode gesehen, weil dies im Gegensatz zu anderen Attribute eine größere Abweichung aufweist und als Wurzelknoten die weiteren Knoten einleitet Bei welchen Änderungen sich die Struktur ändert , kann der **Tabelle 14** entnommen werden. Hierbei ist keine feste Reihenfolge sichtbar, wann sich die Struktur des Entscheidungsbaumes ändert. Der Wurzelknoten wird z.B. bei „minimal gain:0.1; criterion:gain_ration“ von dem Attribut Macine Nitrogen gebildet, sobald sich der Paramter „minimal gain“ auf 0.5 erhöht wird, befindet sich das Attribut in der mittleren Struktur als innerer Knoten. Wie bereits gesehen, haben die Parameter „minimal gain und criterion“ einen Einfluss auf die Aufbaustruktur eines Entscheidungsbaums, jedoch hat die Änderung von „number of folds“ in diesem Fall keine Auswirkung auf die Struktur gehabt.

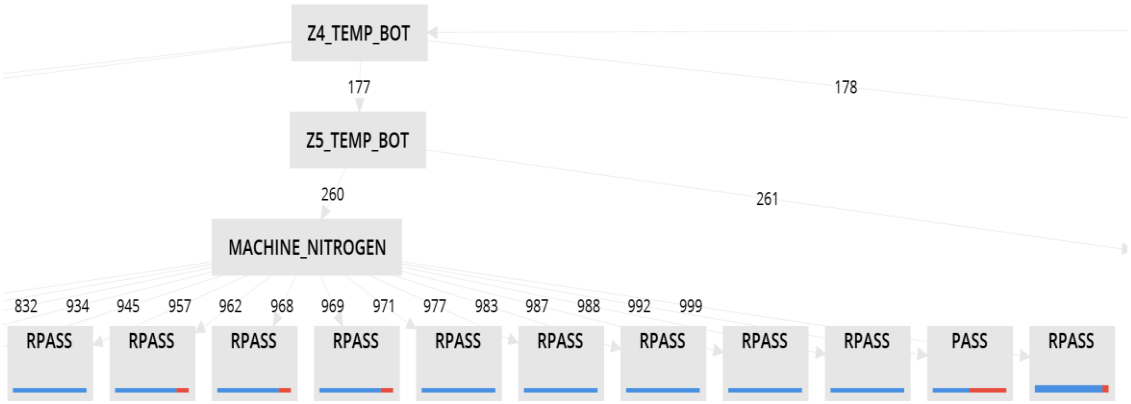


Abbildung 40: Ausschnitt der Veränderten Baumstruktur

Tabelle 14: Änderung der Baumstruktur bei Parametervariationen

Number of folds	Minimal gain	criterion	Baumstruktur	Number of folds	Minimal gain	criterion	Baumstruktur
20	0.1	gain_ration	Machines Nitrogen ist Wurzelknoten	40	0.1	gain_ration	Machines Nitrogen ist Wurzelknoten
20	0.1	Information-gain	Machines Nitrogen ist Wurzelknoten	40	0.1	Information-gain	Machines Nitrogen ist Wurzelknoten

20	0.1	gini_index	Machines Nitrogen kein Wurzelknoten	40	0.1	gini_index	Machines Nitrogen kein Wurzelknoten
20	0.1	accuracy	Machines Nitrogen kein Wurzelknoten	40	0.1	accuracy	Machines Nitrogen kein Wurzelknoten
20	0.5	gain_ration	Machines Nitrogen kein Wurzelknoten	40	0.5	gain_ration	Machines Nitrogen kein Wurzelknoten
20	0.5	Information-gain	Machines Nitrogen kein Wurzelknoten	40	0.5	Information-gain	Machines Nitrogen kein Wurzelknoten
20	0.5	gini_index	Machines Nitrogen ist Wurzelknoten	40	0.5	gini_index	Machines Nitrogen ist Wurzelknoten
20	0.5	accuracy	Machines Nitrogen ist Wurzelknoten	40	0.5	accuracy	Machines Nitrogen ist Wurzelknoten

Untersuchung der Validierungsergebnisse anhand unterschiedlicher Parameter

In diesem Abschnitt wird der ID3-Entscheidungsbaum mit unterschiedlichen Parameter untersucht. Das Validierungsergebnis des Modells bietet eine gute Grundlage zum Vergleich der Ergebnisse mit verschiedenen Parametern. Hierbei wurden folgende Parameter betrachtet: „criterion: gain_ratio, information_gain, gini_index und accuracy; minimal gain und number of folds“. Der Parameter „minimal gain wird auf 0.1, 0.5 und 1 festgelegt. Bei “Number of folds” handelt es sich um einen Parameter der Validierung und gibt die Anzahl der Gruppendatensätze an. In **Tabelle 15** sind die Validierungsergebnisse, die mit den unterschiedlichen Parametereinstellungen erzielten wurden, eingetragen. Hierbei ist die Tabelle in vier Blöcke unterteilt und werden den vier Kriterien des ID3-Entscheidungsbaums zugeordnet. Der erste Block wird dem „criterion: gain_ratio“ zugeordnet. Die horizontale Zeile stellt den „minimal gain“ und die vertikale Zeile die „number of folds“ dar. Die Änderung des „minimal gain“ Parameters ändert das Validierungsergebnis nicht und bleibt bei allen Variationen konstant. Jedoch die Änderung der Anzahl an Gruppendatensätze (number of folds) ändert die Genauigkeit des Modells minimal. Der Wert ändert sich von 10 auf 60 Gruppendatensätze in diesem Fall nur um 0.19%. Die Anzahl der Gruppendatensätze wird jeweils um 20 erhöht und somit die Änderung der Validierung beobachtet. Die Minimale Steigung der Genauigkeit des Modells ist so minimal,

dass die Entwicklung des Modells kein besonderes Gewicht hat. Der Durchschnittswert der Validierung überschreitet nicht die 60% trotz Parametervariationen. Des Weiteren werden die weiteren Blöcke betrachtet und festgestellt, dass die Änderung von „criterion“ die Genauigkeit des Modells anhand der vorliegenden Datenmenge nicht verändert. Das Validierungsergebnis wird anhand der Parameter „minimal gain und criterion“ nicht beeinflusst.

Tabelle 15: Validierungsergebnis mit unterschiedlichen Parameter

	gain_ratio			**information_gain**		
Number of folds	minimal gain			minimal gain		
	0.1	0.5	1.0	0.1	0.5	1.0
10	60.52%	60.52%	60.52%	60.52%	60.52%	60.52%
20	60.54%	60.54%	60.54%	60.54%	60.54%	60.54%
40	60.65%	60.65%	60.65%	60.65%	60.65%	60.65%
60	60.71%	60.71%	60.71%	60.71%	60.71%	60.71%
80	60.68%	60.68%	60.68%	60.68%	60.68%	60.68%
100	60.78%	60.78%	60.78%	60.78%	60.78%	60.78%
	gini_index			**accuracy**		
Number of folds	minimal gain			minimal gain		
	0.1	0.5	1.0	0.1	0.5	1.0
10	60.52%	60.52%	60.52%	60.52%	60.52%	60.52%
20	60.54%	60.54%	60.54%	60.54%	60.54%	60.54%
40	60.65%	60.65%	60.65%	60.65%	60.65%	60.65%
60	60.71%	60.71%	60.71%	60.71%	60.71%	60.71%
80	60.68%	60.68%	60.68%	60.68%	60.68%	60.68%
100	60.78%	60.78%	60.78%	60.78%	60.78%	60.78%

Der Entscheidungsbaum mit dem ID3-Algorithmus hat den Vorteil, dass die Analyse durch die Festlegung eines Zielattributes gestartet werden kann. Speziell in Bezug auf Prozessparameter und Bauteilqualität spielt die Ausprägung wie Pass, Rpass oder Repair eine wichtige Rolle um die Prozessparameter zu identifizieren, die auch mängelfreie Produkte produzieren. Es bietet sich an, sich die Baumstrukturen anzugucken und entsprechende Maßnahmen zu ergreifen oder mithilfe des trainierten Modells eine Voraussage bezüglich der Qualität des Produktes zu treffen. In diesem Fall ist weder die Baumstruktur noch das trainierte Modell zufriedenstellend um Aussagen zu treffen. Dies kann an dem Parameter „Machine Nitrogen“ liegen, weil die Anzahl

an Abweichungen recht hoch ist und somit die Baumstruktur als auch das Modell stark beeinflussen. Würden die Prozessparameter in überschaubaren Rahmen liegen, würde der Entscheidungsbaum mit der Relation zum Qualitätsstatus eine gute Möglichkeit bieten, Prozesse zu untersuchen und zu optimieren.

Darstellung und Bewertung der Ergebnisse der k-Means Clusteranalyse

In diesem Abschnitt werden die Ergebnisse, die mit der k-Means Clustermodellierung durchgeführt wurde, dargestellt und bewertet.

Clusteranalyse mit Erhöhung der Anzahl der Cluster k

In dieser Arbeit wird mit unterschiedlicher Anzahl an Clustern *k* und Iterationsverläufen *max runs* gearbeitet. Die Anzahl an Clustern kann unter *Paramters; k* individuell angepasst. Die Anzahl der Iterationsverläufe wird unter *max runs* eingestellt. Hierbei erfolgt die Abstandsberechnung unter der Standardeinstellung *measure types: MixedMeasures; mixed measure: MixedEuclideanDistance.*

Die Clusteranalyse wird ohne Einschränkung an allen Prozessparametern durchgeführt mit unterschiedlicher Anzahl an Clustern und Iterationsverläufen. Zuerst wird *k* stetig verändert und Anzahl der Iterationsverläufe konstant gehalten. In **Tabelle 16** werden die Experimentversuche mit unterschiedlicher Anzahl an Clustern *k* und die dazugehörigen Ergebnisse dargestellt. Die Iteration *max runs* sind bei allen ausgewählten *k* unverändert und liegt bei 10 Verläufe. Um den Unterschied zwischen den verschiedenen Clusteranzahlen darzustellen, wird der Performancewert betrachtet. Hierbei werden zwei Werte besonders betrachtet: Average within cluster distance (AV) und der Davies Bouldin Index (DBI). Beide Werte spiegeln die Qualität des Clusters wieder. Zur Vereinfachung wird für den weiteren Verlauf auf ein Validierungswert festgelegt, in dem Fall auf Average within cluster distance (AV). Den Davies Bouldin Index (DBI) wurde zur Vervollständigung zusätzlich in der Tabelle angegeben und kann entsprechend wie AV interpretiert werden. In diesem Experiment wurde die Anzahl von 2 auf 60 Clustern erhöht. Die Erhöhung von „k = 61“ hat keine Bedeutung mehr gehabt, weil die maximale Anzahl an Clustern für die vorliegende Datenmenge erreicht wurde. Jede weitere Erhöhung von *k* würde Clustern hervorbringen, die keine Daten beinhalten und nicht sinnvoll sind. Mit zunehmender Steigerung der Clusteranzahl wird die Qualität der Cluster stetig verbessert, der Abstand der Datenpunkte zum Centroiden verringert sich. In der Tabelle wird sichtbar, dass der Validierungswert AV sich mit steigendem *k* verringert. Je mehr Cluster existieren, desto strukturierter werden die Daten zugeordnet. Die Datenmenge wird aufgespalten und wird zunehmend kleiner. Die Anzahl der Datenmenge variiert je Cluster, je kleiner die Menge desto optimaler auch der Abstand. Bei „k = 2, 3, 4“ werden große Datenmengen „willkürlich“ den vorhandenen Clustern zugeordnet und dementsprechend ist der Abstand „gross“ und nicht

optimal. Die Erhöhung von „*k auf 20, 30"* bewirkt, dass die Rechenlaufzeit für die Bildung der Cluster entsprechend höher liegt als bei „k = 2, 3".

Tabelle 16: Analyse mit unterschiedlicher Anzahl an Clustern

Anzahl der Cluster *k*	**Anzahl der Iterationsverläufe *max runs***	**Durchschnittlicher Abstand zum Centroiden**
2	10	AV: -2,463 DBI: -1,842
3	10	AV: -2,190 DBI: -1,862
4	10	AV: -1,939 DBI: -1,818
5	10	AV: -1,848 DBI: -1,743
6	10	AV: -1,785 DBI: -1,711
7	10	AV: -1,644 DBI: -1,754
8	10	AV: -1,624 DBI: -1,625
9	10	AV: -1,583 DBI: -1,700
10	10	AV: -1,616 DBI: -1,716
15	10	AV: -1,486 DBI: -1,644
20	10	AV: -1,387 DBI: -1,524
30	10	AV: -1293 DBI: -
40	10	AV: -1,214 DBI: -1,410
60	10	AV: -1,144 DBI: -

Clusteranalyse mit Erhöhung der Iterationsabläufe bei gleichbleibender Anzahl an Cluster

Nach dem die Auswirkung der Clusteranzahl auf den AV untersucht wurde, wird in diesem Abschnitt der Parameter max runs erhöht und die Auswirkung auf den Validierungswert AV überprüft. Hierbei wird „max runs auf 20, 30 ,40“ Iterationsverläufen erhöht. In **Tabelle 17** wird der Validierungswert AV zur Anzahl *k* und den unterschiedlichen Iterationsverläufen aufgelistet. Zum Vergleich wird der Standardparameter „max runs = 10“ zusätzlich zu den weiteren Iterationsabläufen in der Tabelle miteingefügt, um die AV-Änderungen besser nachzuvollziehen. Die Auswirkung des Parameters *max runs* auf den AV bei der Clusteranzahl „k=2, 3, 4, 5“ ist erstmal sehr gering und auch erst ab 30 Abläufen sichtbar. Hierbei ist eine minimale Verbesserung der Qualität des Clusters feststellbar. Erst ab „k = 10“ wird die Distanz mit der erhöhten Iteration kleiner, dies beginnt bereits bei „max runs = 20“. Die Änderungen des AV befindet sich größtenteils bei 20 und 30 Iterationsverläufen, zwischen 30 und 40 sind soweit keine Veränderungen zu sehen. Anhand der Tabelle wird deutlich, dass die Bildung von „k=1 - 9“ Clustern anhand von 10 Iterationsverläufen erfolgen kann, ohne das weitere Verbesserungen stattfinden. Bei einer kleinen Anzahl an Clustern reicht in diesem Fall der Standardparameter „max runs=10“ für die Bildung von Clustern aus, ohne das weitere Optimierungen stattfinden. Je größer die Anzahl der Cluster wird, desto mehr Iterationen werden benötigt, um das gewünschte Ziel zu erreichen. Die Qualität der Cluster wird durch die Erhöhung verbessert, in dem die Clusterbildung weiter optimiert um somit der Abstand zum Centroiden gering werden.

Tabelle 17: Auswirkung der Iterationsverläufe auf AV

Anzahl der Cluster *k*	**Anzahl der Iterationsverläufe *max runs***			
	10	20	30	40
2	AV: -2,463 DB: -1,842	AV: -2,463 DBI:- 1,843	AV: -2,463 DBI:- 1,843	AV: -2,463 DBI:- 1,843
3	AV: -2,190 DB: -1,862	AV: -2,190 DBI: -1,862	AV: -2,190 DBI: -1,862	AV: -2,190 DBI: -1,862
4	AV: -1,939 DB: -1,818	AV: -1,939 DBI: -1,818	AV: -1,937 DBI: -1,802	AV: -1,937 DBI: -1,802
5	AV: -1,848 DB: -1,743	AV: -1,848 DBI: -1,904	AV: -1,844 DBI: -1,761	AV: -1,844 DBI: -1,761
10	AV: -1,616 DB: -1,716	AV: -1,536 DBI: -1,722	AV: -1,498 DBI: -1,637	AV: -1,498 DBI: -1,637
20	AV: -1,387 DB: -1,524	AV: -1,359 DBI: -1,451	AV: -1,319 DBI: -1,499	AV: -1,319 DBI: -1,499

30	AV: -1293 DB: -	AV: -1,244 DBI: -1,377	AV: -1,244 DBI: -1,377	AV: -1,244 DBI: -1,377
40	AV: -1,214 DB: -1,410	AV: -1,214 DBI: -1,410	AV: -1,187 DBI: -	AV: -1,187 DBI: -
60	AV: -1,144 DB: -	AV: -1,132 DB: -	AV: -1,132 DB:	AV: -1,132 DB:

Vergleich der Centroiden Table mit unterschiedlichem Clusteranzahl

Bislang wurden die Auswirkung der verschiedenen Parameter anhand von „Average within cluster distance (AV)“ betrachtet, wie sich der Abstand zum Centroiden verändert und somit die Qualität des Clusters beeinflusst. In diesem Abschnitt wird untersucht wie die Änderung der Parameter sich auf die Daten wiederspiegelt. Bislang kann anhand der letzten Untersuchungen ausgegangen werden, dass eine Optimierung bezüglich der Clusterstruktur erfolgt. Dies wird anhand des Validierungswert AV festgemacht. Ob die Optimierung auch in der Datenmenge wiederspiegelt, wird anhand der Centroiden Table und Cluster Plot geklärt. Mit Hilfe des Centroid Table kann eingesehen werden, welche Attributwerte das jeweilige Cluster dominieren. Je geringer der Wert, desto geringer die Relevanz in dem jeweiligen Cluster. Das Cluster Plot stellt die Attributwerte mit der größten Relevanz graphisch dar. Für die Untersuchung werden zwei Centroiden Table mit dem jeweiligen Cluster Plot dargestellt. Die Parametereinstellung sind in diesem Fall „k=2, max runs=10“ und „k=20, max runs=10“ und werden unter der jeweiligen Abbildung erwähnt.

In **Abbildung 41** ist ein Ausschnitt des Centroid-Table mit den Parametereinstellung „k = 2, max runs = 10“ dargestellt. Es befinden sich drei Spalten und beinhaltet Attributwerte und die zwei Cluster. Auffällig in der Abbildung sind stark dominierende als auch geringe Attributwerte, die in den beiden Clustern enthalten sind. Die Vielfalt ist in den Clustern groß, es gibt Werte die kleine (z.B. Z1_TEMP_TOP=195, 0,999) als auch große Abstände (MACHINE NITROGEN=1006, 0,015) zum Centroiden bilden. Es werden Cluster gebildet, aber die Aussagekraft über die Attributwerte ist nicht eindeutig und zu unübersichtlich. Mithilfe des Cluster-Plot (s. **Abbildung 42**) können die dominierenden Werte graphisch abgelesen werden. Die Attributwerte, die keine besondere Rolle in den Clustern haben, sind auf der Achse im Bereich des Nullpunktes zu finden. Die Differenzierung der Cluster erfolgt farblich.

Attribute	cluster_0 ↓	cluster_1
Z1_TEMP_TOP = 195	0.999	0.997
Z3_TEMP_TOP = 195	0.999	0.997
Z3_TEMP_BOT = 169	0.997	0
Z4_TEMP_TOP = 200	0.997	1
Z6_TEMP_BOT = 260	0.991	0.998
Z2_TEMP_BOT = 146	0.937	0.710
Z5_TEMP_BOT = 260	0.935	0.955
Z7_TEMP_TOP = 80	0.934	0.963
Z1_TEMP_BOT = 133	0.668	0.177
Z4_TEMP_BOT = 178	0.569	0.864
Z4_TEMP_BOT = 179	0.384	0.003
Z1_TEMP_BOT = 134	0.168	0
Z1_TEMP_BOT = 132	0.164	0.469
Z7_TEMP_TOP = 81	0.066	0.037
Z2_TEMP_BOT = 147	0.063	0
Z4_TEMP_BOT = 177	0.046	0.105
Z5_TEMP_BOT = 261	0.034	0.016
Z5_TEMP_BOT = 259	0.028	0.029
MACHINE_NITROGEN = 990	0.017	0.013
MACHINE_NITROGEN = 1006	0.015	0.013
MACHINE_NITROGEN = 1010	0.015	0.013
MACHINE_NITROGEN = 1018	0.015	0.010

Abbildung 41: Auschnitt Centroid-Table für k=2, max runs=10

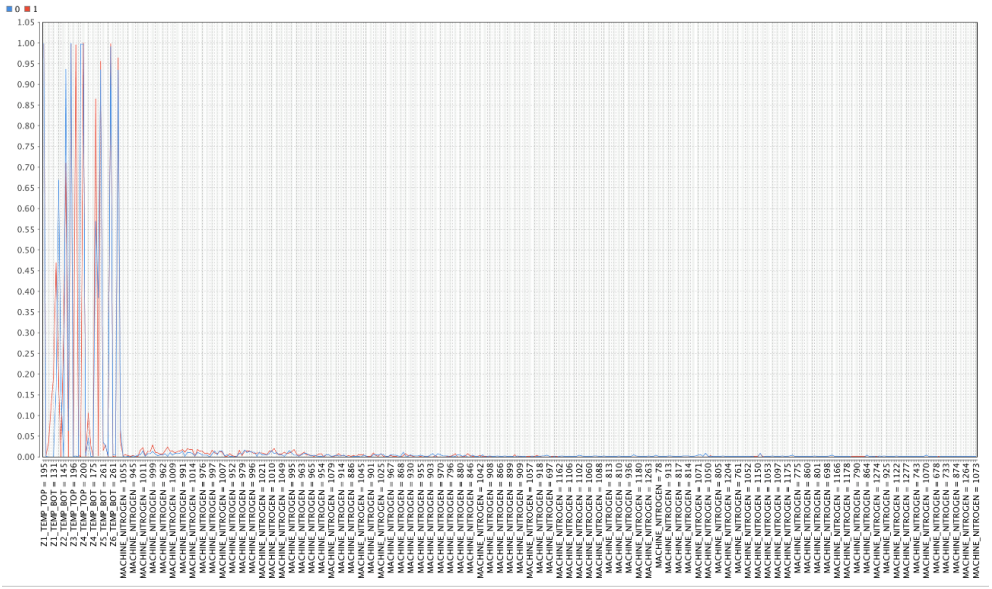


Abbildung 42: Cluster-Plot für k=2, max runs=10

Nach der Betrachtung von „k = 2, max runs = 10“ wird die Anzahl der Cluster auf „k = 20, max runs = 10“ erhöht um somit die Clusterstruktur zu analysieren und den Unterschied darzustellen. Die vorliegende Datenmenge wird nicht wie im vorherigen Fall auf zwei Cluster, sondern auf 20 Cluster verteilt, hierbei entstehen mehr Möglichkeiten strukturiert zu clustern. Es werden Datenwerte nicht „willkürlich“ in Clustern untergebracht. In **Abbildung 43** wird verdeutlicht, dass nur Datenwerte einem Cluster zugeordnet wurden, die in einer Kombination Sinn ergeben. In dieser Arbeit handelt es sich konkret um Prozessparameter, die Cluster können mögliche

Kombinationen von Prozessparametern angeben. In der Abbildung unter „cluster_3“ sieht man ein gutes Beispiel, wo nicht jeder Wert, die in der Liste vorhanden sind, im Cluster auftauchen. In dem Fall nur Werte, wo kein Abstand zum Centroiden besteht und prozesstechnisch Sinn macht. Bei k = 2 besteht das Problem, dass jedes Attributwert in den Clustern auftaucht, dieses Problem besteht bei einer Erhöhten Anzahl an Cluster nicht. Die „willkürliche“ Zuordnung erfolgt nicht, sondern nur die Werte die auch zur der Clusterstruktur passen, wird „aufgenommen“. Der dazugehörige Cluster-Plot (s. **Abbildung 44**) zeigt auch hier graphisch die dominierenden Werte, jedoch besteht der Unterschied zum letzten Plot, dass Werte vom Attribut Maschine Nitrogen, die aufgrund deren Vielzahl zuvor keine besondere Rolle spielen, nun vereinzelnde Werte dominieren.

Attribute	cluster...	cluster...	cluster...	cluster_3 ↓	cluster...	cluster...	cluster...	cluster...	cluster...	cluster...	cluster...	cluster...
Z1_TEMP_TOP = 195	1	1	1	1	1	1	0.996	1	1	1	1	1
Z1_TEMP_BOT = 133	1	1	0	1	0	1	0	0.250	0.333	1	0.636	0
Z2_TEMP_BOT = 146	1	1	1	1	0	1	1	0.875	1	1	0.818	0
Z3_TEMP_TOP = 195	1	1	1	1	1	1	0.996	1	1	1	1	1
Z3_TEMP_BOT = 169	1	1	1	1	0	1	0	0	0	1	1	1
Z4_TEMP_TOP = 200	1	1	1	1	1	0.923	1	1	1	1	1	1
Z4_TEMP_BOT = 178	0.948	0	0.980	1	0	1	0.988	0.938	1	0	0	0.700
Z5_TEMP_BOT = 260	1	0.946	0.969	1	0.969	0	0.935	1	1	1	0.909	0.900
Z6_TEMP_BOT = 260	0.994	0.987	0.990	1	1	1	0.996	1	1	1	0.909	1
Z7_TEMP_TOP = 80	0.948	1	0.929	1	0.969	0.846	1	0	1	1	0	1
MACHINE_NITROGEN = 967	0	0	0.020	1	0	0	0	0	0	0	0	0
Z1_TEMP_TOP = 196	0	0	0	0	0	0	0.004	0	0	0	0	0
Z1_TEMP_BOT = 129	0	0	0	0	0.313	0	0	0	0	0	0	0
Z1_TEMP_BOT = 130	0	0	0	0	0.608	0	0	0	0	0	0	0
Z1_TEMP_BOT = 131	0	0	0	0	0.047	0	0	0.062	0	0	0	0
Z1_TEMP_BOT = 132	0	0	1	0	0.031	0	1	0.688	0.667	0	0.091	0
Z1_TEMP_BOT = 134	0	0	0	0	0	0	0	0	0	0	0.273	1
Z2_TEMP_BOT = 145	0	0	0	0	1	0	0	0.125	0	0	0	0
Z2_TEMP_BOT = 147	0	0	0	0	0	0	0	0	0	0	0.182	1
Z3_TEMP_TOP = 196	0	0	0	0	0	0	0.004	0	0	0	0	0
Z3_TEMP_BOT = 168	0	0	0	0	0.953	0	1	1	1	0	0	0
Z3_TEMP_BOT = 167	0	0	0	0	0.047	0	0	0	0	0	0	0

Abbildung 43: Centroid-Table für k=20, max runs=10

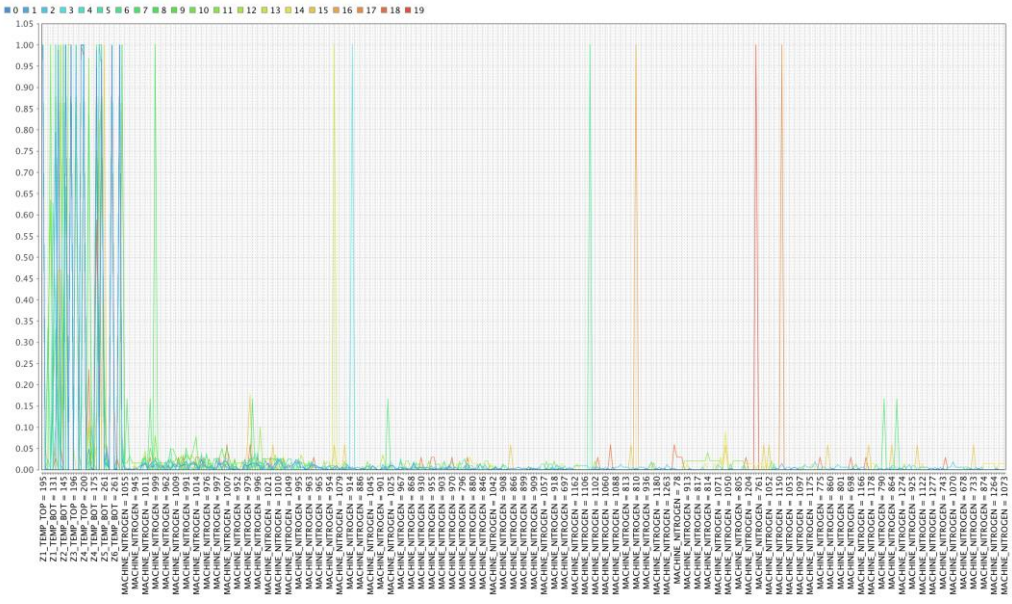


Abbildung 44: Cluster-Plot für k=40, max runs=10

Die Clusteranalyse bietet mit den bisherigen Erkenntnissen eine gute Möglichkeit große Datenmengen zu verarbeiten und Informationen gewinnen. Dies ist jedoch stark mit der Clusteranzahl gekoppelt. Anhand von zwei Clustern ist es recht schwierig in großen Datenmengen eine Struktur reinzubringen. Die Daten müssen gezerrt werden, dafür ist die Bildung von mehreren Clustern notwendig. Die Anzahl an Clustern muss der Datenmenge angemessen sein, somit verringern sich die Abstände zum Centroiden. Die Iterationsverläufe verändern minimal den Abstand, dies ist jedoch nicht so stark an den Datenclustern sichtbar, wie dies bei der Veränderung des Parameters *k* deutlich wird. In Bezug auf Prozessparametern bietet das k-Means Clusterverfahren gute Möglichkeiten Kombinationen und Ausreißer zu erkennen, dies ist mithilfe des Centroid-Table als auch mit dem Cluster-Plot gut möglich.

Darstellung und Bewertung der Ergebnisse FP-Growth Assoziationsanalyse

Zum Schluss wird in diesem Abschnitt die Ergebnisse der FP-Growth Assoziationsanalyse betrachtet und bewertet.

Untersuchung der Ergebnismenge anhand der Parameter „min number of itemsets“ und „min support“

In diesem Abschnitt wird anhand der Parameter „min number of itemsets und min support“ der Zusammenhang aufgezeigt, wie sich diese Einstellungen auf die Anzahl der Ergebnismenge und Anzahl der Items auswirkt. Der Parameter „min support“ wurde hierbei auf 0.1, 0.5 und 0.9 festgelegt und „min number of itemset“ von 50 auf 800 gesteigert. Die maximale Anzahl der Ergebnismenge wird bei „min support = 0.1“ ausgegeben und die Anzahl der Items steigert sich entsprechend auf sieben Spalten. In diesem Fall wird bei einem „min support = 0.1“ laut Tabelle bis zu sieben Items angezeigt. Die Steigerung des Wertes von 0.1 auf 0.5, 0.9 führte zu einer Eingrenzung der Ergebnismenge und Anzahl der Items. Des Weiteren kann mithilfe des Parameters „min number of itemsets“ die maximale Anzahl der Itemsets festgelegt werden, die erstellt werden sollen. Die Erhöhung der Anzahl bewirkt entsprechend eine Erhöhung der Ergebnismenge, die wiederum mit dem „min support“ reguliert werden kann. Je kleiner der Parameter „min support“ festgelegt wird, desto mehr an Ergebnismengen werden ausgegeben, jedoch ist die Relevanz und Aussagekraft als zu niedrig einzustufen.

Wie sich die Variation der Parameter auf die Ergebnismenge auswirkt, kann der **Tabelle 18** entnommen werden.

Tabelle 18: Bestimmung der Ergebnismenge anhand von Paramtervariationen

min number of itemsets	min support	Anzahl an Sets	max. Anzahl an Items
50	0.1	344	7
50	0.5	59	5
50	0.9	51	5
100	0.1	344	7
100	0.5	123	5
100	0.9	115	5
150	0.1	344	7
150	0.5	155	6
150	0.9	167	6
250	0.1	344	7
250	0.5	255	7
250	0.9	167	6
450	0.1	467	7
450	0.5	305	7
450	0.9	167	6
800	0.1	808	7
800	0.5	305	7
800	0.9	167	6

In **Abbildung 45** wird die Ergebnistabelle vom FP-Growth Algorithmus vorgestellt. Die Tabelle gilt für folgende Parametereinstellung „min number of itemsets = 100 und min support = 0.5“. Die maximale Anzahl der Ergebnismenge beträgt 123 mit einer maximalen Anzahl von 5 Items. Den höchsten Support erhielt das Attribut Z6_TEMP_BOT= 260 mit einem Supportwert von 0.994. Die maximale Anzahl an Items wird hier zwar mit 5 angegeben, aber nicht jede Ergebnismenge beinhaltet automatisch diese Menge. Die 5 Itemspalten sind bei einem hohen Supportwert von 0.8-1 so gut wie gar nicht ausgefüllt. Dies erfolgt erst unter einem Supportwert von 0.4, aber auch nur vereinzelnd. Die Prozessparameter mit starken Abweichungen wie Machine Nitrogen treten in der Ergebnismenge der Tabelle nicht auf. Wie zuvor erwähnt, hat konkret dieser Prozessparameter eine starke Variation in den Werten und das Verhältnis zu den Gesamtdaten ist zu gering um einen bestimmten Supportwert zu erreichen. In der Ergebnismenge finden sich eher Prozessparameter wieder, die eine kleine Abweichungen vorweisen.

No. of Sets: 123
Total Max. Size: 5

Min. Size: 1
Max. Size: 5
Contains Item:

Update View

Size	Support	Item 1	Item 2	Item 3	Item 4	Item 5
1	0.994	Z6_TEMP_BOT = 260				
1	0.944	Z5_TEMP_BOT = 260				
2	0.940	Z6_TEMP_BOT = 260	Z5_TEMP_BOT = 260			
1	0.836	Z2_TEMP_BOT = 146				
2	0.831	Z6_TEMP_BOT = 260	Z2_TEMP_BOT = 146			
2	0.788	Z5_TEMP_BOT = 260	Z2_TEMP_BOT = 146			
3	0.784	Z6_TEMP_BOT = 260	Z5_TEMP_BOT = 260	Z2_TEMP_BOT = 146		
1	0.700	Z4_TEMP_BOT = 178				
2	0.697	Z6_TEMP_BOT = 260	Z4_TEMP_BOT = 178			
2	0.669	Z5_TEMP_BOT = 260	Z4_TEMP_BOT = 178			
3	0.667	Z6_TEMP_BOT = 260	Z5_TEMP_BOT = 260	Z4_TEMP_BOT = 178		
1	0.642	Qualitätsstatus = R...				
2	0.640	Z6_TEMP_BOT = 260	Qualitätsstatus = R...			
2	0.619	Z2_TEMP_BOT = 146	Z4_TEMP_BOT = 178			
3	0.616	Z6_TEMP_BOT = 260	Z2_TEMP_BOT = 146	Z4_TEMP_BOT = 178		
2	0.608	Z5_TEMP_BOT = 260	Qualitätsstatus = R...			
3	0.606	Z6_TEMP_BOT = 260	Z5_TEMP_BOT = 260	Qualitätsstatus = R...		
3	0.591	Z5_TEMP_BOT = 260	Z2_TEMP_BOT = 146	Z4_TEMP_BOT = 178		
4	0.590	Z6_TEMP_BOT = 260	Z5_TEMP_BOT = 260	Z2_TEMP_BOT = 146	Z4_TEMP_BOT = 178	

Abbildung 45: FP-Growth Analyse Ergebnis

Untersuchung der Abhängigkeitsregeln mit den Parametern „min support und min confidence“

Neben der Gruppierung von Datensätzen zu Itemsets erarbeiten Assoziationsverfahren mit Hilfe des *Create Association Rules Operators* Regeln. Es werden Abhängigkeitsregeln mit den Prozessparametern formuliert. Es werden Prozessparameterkombinationen vorgeschlagen, die anhand von Support und Konfidenz bestimmt werden. Der Operator für die Erstellung der Regeln arbeitet mit verschiedenen Kriterien: confidence, lift, conviction, ps, gain und laplace. Bei „confidence“ muss der Parameter „min confidence“ und bei den anderen Kriterien „min criterion value“ mit entsprechenden Wert ausgefüllt werden. Je nach dem welches Kriterium hier ausgewählt wird, die Ergebnisse erscheinen alle in einer Datentabelle. Da der „confidence“ Wert bereits bekannt ist, wird damit weitergearbeitet. Bei der Ausführung des Modells werden bereits alle möglichen Regeln erfasst, jedoch ist hier entscheidend wie hoch der „min confidence“ Wert ausgewählt wird. Für die vorliegende Datenmenge wurden 5252 Regeln entwickelt. Je nach wie „min support und min confidence“ gewählt werden, filtert er die entsprechenden Regeln raus und somit verringert bzw. erhöht sich die Anzahl der Regeln. Um verlässlichen Regeln zu erhalten, sollten beide Parameter in dem Bereich 0.8 - 0.9 liegen. Damit die entwickelten Regeln eine Aussagekraft haben, sollten die beiden Kriterien hoch angesetzt werden. In **Abbildung 46** ist ein Ausschnitt über die abgeleiteten Regeln mit den Parametern „min support = 0.9 und min

confidence = 0.9“ dargestellt. Die Anzahl der Regeln hat sich 5252 auf 20 verringert. Hierbei ist anzumerken, dass nur ein Parameter, in diesem Fall „Confidence“ die min. Vorgabe erfüllt.

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s
1	Z1_TEMP_BOT = 133	Z2_TEMP_BOT = 146	0.450	1	1	-0.450	0.074
2	Z5_TEMP_BOT = 260, Z3_TEMP_BOT = 168	Z6_TEMP_BOT = 260	0.423	1	1	-0.423	0.002
3	Z5_TEMP_BOT = 260, Z1_TEMP_BOT = 132	Z6_TEMP_BOT = 260	0.284	1	1	-0.284	0.002
4	Z6_TEMP_BOT = 260, Z1_TEMP_BOT = 133	Z2_TEMP_BOT = 146	0.447	1	1	-0.447	0.073
5	Z5_TEMP_BOT = 260, Z1_TEMP_BOT = 133	Z2_TEMP_BOT = 146	0.428	1	1	-0.428	0.070
6	Z4_TEMP_BOT = 178, Z1_TEMP_BOT = 133	Z2_TEMP_BOT = 146	0.308	1	1	-0.308	0.051
7	Qualitätsstatus = RPASS, Z1_TEMP_BOT = 133	Z2_TEMP_BOT = 146	0.286	1	1	-0.286	0.047
8	Z3_TEMP_BOT = 169, Z1_TEMP_BOT = 133	Z2_TEMP_BOT = 146	0.370	1	1	-0.370	0.061
9	Z6_TEMP_BOT = 260, Z5_TEMP_BOT = 260, ...	Z2_TEMP_BOT = 146	0.424	1	1	-0.424	0.070
10	Z5_TEMP_BOT = 260, Z2_TEMP_BOT = 146, ...	Z6_TEMP_BOT = 260	0.300	1	1	-0.300	0.002
11	Z5_TEMP_BOT = 260, Z4_TEMP_BOT = 178, ...	Z6_TEMP_BOT = 260	0.365	1	1	-0.365	0.002
12	Z5_TEMP_BOT = 260, Qualitätsstatus = RPASS...	Z6_TEMP_BOT = 260	0.293	1	1	-0.293	0.002
13	Z6_TEMP_BOT = 260, Z4_TEMP_BOT = 178, ...	Z2_TEMP_BOT = 146	0.306	1	1	-0.306	0.050
14	Z6_TEMP_BOT = 260, Qualitätsstatus = RPASS...	Z2_TEMP_BOT = 146	0.285	1	1	-0.285	0.047
15	Z6_TEMP_BOT = 260, Z3_TEMP_BOT = 169, ...	Z2_TEMP_BOT = 146	0.367	1	1	-0.367	0.060
16	Z5_TEMP_BOT = 260, Z4_TEMP_BOT = 178, ...	Z2_TEMP_BOT = 146	0.297	1	1	-0.297	0.049
17	Z5_TEMP_BOT = 260, Z3_TEMP_BOT = 169, ...	Z2_TEMP_BOT = 146	0.350	1	1	-0.350	0.057
18	Z6_TEMP_BOT = 260, Z5_TEMP_BOT = 260, ...	Z2_TEMP_BOT = 146	0.295	1	1	-0.295	0.048
19	Z5_TEMP_BOT = 260, Z2_TEMP_BOT = 146, ...	Z6_TEMP_BOT = 260	0.292	1	1	-0.292	0.002
20	Z6_TEMP_BOT = 260, Z5_TEMP_BOT = 260, ...	Z2_TEMP_BOT = 146	0.346	1	1	-0.346	0.057

Abbildung 46: Ansicht der Assoziationsregeln

Empfehlung geeigneter Data-Mining Methode für die Elektronikfertigung

Die Modellierung für drei verschiedene Methoden ist durch RapidMiner realisiert worden. Die Modelle wurden anhand der vorliegenden Datensätze trainiert und getestet und für die Methoden übliche Ergebnisse ausgegeben. Um eine Auswahl zwischen den eingesetzten Methoden zu treffen, die für Produktionsdaten geeignet sind, wurden die jeweiligen Modellierungen mit den vorhandenen Daten getestet und Ergebnisse bewertet.

1) ID3-Entscheidungsbaum

   Der Vorteil beim Entscheidungsbaum besteht darin, dass ein Label-Attribut definiert wird, die praktisch als Zielattribut gilt. Die Baumstruktur verfolgt das Ziel dieses Zielattribut mit den vorliegenden Daten zu erfüllen. Auf diese Arbeit bezogen, wird die Qualität eines Bauteils mit den entsprechenden Prozessparametern in Relation gesetzt um somit gegebenenfalls die Parameterwerte zu optimieren. Daher kann bei einem Entscheidungsbaum das Attribut Qualität als Zielattribut definiert werden und entsprechend die Strukturen bis zum Blattknoten verfolgen. Das Ergebnis wird visuell dargestellt und somit einfacher nachzuvollziehen. Des Weiteren kann das Modell mit Testdaten getestet werde und somit eine Voraussage über das Zielattribut machen. Der Nachteil bestand jedoch darin, falls ein Parameter eine unübliche Abweichung hat und

nicht minimal vom Sollwert abweicht, wird die Baumstruktur komplex aufgebaut und wird unübersichtlich. Die Abweichung des Parameters Machine Nitrogen hat die Entscheidungskraft des Entscheidungsbaumes in dieser Arbeit drastisch eingeschränkt. Sobald die Parameter keine hohe Abweichung haben, dient der Entscheidungsbaum als eine gute Analysemethode.

2) k-Means Clusterverfahren

   Für das k-Means Clusterverfahren wird kein Zielattribut wie beim Entscheidungsbaum definiert. Die Aussagekraft des Clusterverfahren wird hauptsächlich von der Anzahl der Cluster bestimmt. Sobald die Anzahl der Cluster an die Datenmenge angepasst wird, ist der Abstand zu den einzelnen Datenpunkten nicht sehr hoch und gibt im Cluster Prozessparameter an. Anhand des Centroid-Table und Cluster-Plot können somit vereinfacht Ausreißer entdeckt werden, die beseitigt werden können. Als Beispiel wäre Machine Nitrogen zu nennen, deren Abstände zum Centroiden aufgrund der vielen unterschiedlichen Prozesswerte sehr hoch war und dementsprechend kaum in einem Cluster dominierte. Das Clusterverfahren kann bei der Auswertung von Produktionsdaten für die Entdeckung von Ausreißer sehr hilfreich sein und z.B. den Entscheidungsbaum unterstützen.

3) FP-Growth Assoziationsverfahren

   Das Assoziationsverfahren hebt bei der Analyse von Produktionsdaten, Kombination von Prozessparametern hervor. Damit kann ein Überblick verschafft werden, welcher Prozesswert häufig und relativ konstant vorkommt. Dieses Verfahren eignet sich generell zum Auffinden von Ausreißern oder von bestimmten Prozessparameterkombinationen. In diesem Fall kann es als Unterstützung zum Entscheidungsbaum ausgeführt werden.

5. Handlungsempfehlung für die Elektronikfertigung des Pumpenherstellers

In diesem vorletzten Kapitel wird zur Einordnung dieser Arbeit das Unternehmen WILO SE vorgestellt und Handlungsempfehlungen vorgeschlagen.

5.1 Vorstellung Wilo SE

Diese Arbeit wurde mit der Unterstützung der Wilo SE verfasst, daher wird im Folgenden das Unternehmen kurz vorgestellt. Die Wilo SE ist einer der weltweit führenden Hersteller im Hightech-Pumpen-Bereich für Heizungs-, Kälte- und Klimatechnik, die Wasserversorgung sowie die Abwasserentsorgung. Das Unternehmen ist in den Marktsegmenten Building Services, Water Management und Industry tätig und hat ihr Portfolio aus Produkten, Systemlösungen und Serviceleistungen konsequent und maßgeschneidert auf den jeweiligen Bedarf der Kunden in diesen drei Marktsegmenten ausgerichtet.

Die Wilo SE produziert Pumpen und Pumpensysteme dezentral an 16 Standorten in Europa, Asien und Amerika. In Deutschland hat die Wilo SE vier Produktionsstandorte in Dortmund, Hof, Minden und Oschersleben. Die Leiterplatinen für die Pumpenmodule entwickelt und produziert das Unternehmen in Oestrich unabhängig von externen Dienstleister. Die Gruppe Electronic & Motors beliefert weltweit alle Standorte mit der elektronischen Komponente und sichert somit die interne Produktion von Wasserförderpumpen. Die Wilo SE verfolgt zur Zeit einer ihrer wichtigsten Ziele, die Digitalisierung 2020. Dieses Ziel soll sich nicht nur an dem Produkt, erste smarte Pumpe, umgesetzt werden, sondern auch intern im Unternehmen. Im Zuge der digitalen Transformation und Industrie 4.0 werden Produktionsprozess neu konzipiert und ermöglichen in der Produktion mithilfe von vernetzten Maschinen und Produkten eine Echtzeitprüfung der Prozessdaten.

5.2 Empfehlung für die Verbesserung der Analysequalität

Die Grundlage für die folgenden Handlungsempfehlungen ergeben sich aus den gewonnen Erfahrungen und Erkenntnissen, die während der Zeit bei Wilo SE und Analyse der Firmendaten gewonnen wurden. Die Empfehlungen gelten nicht nur für den untersuchten Bereich, diese können auf weitere Anlagen und Systeme ausgeweitet werden.

5.2.1 Aufbau der Datenstruktur in MES

Um eine Analyse schnell und effektiv an Prozessdaten auszuführen, müssen exportierte Daten in der richtigen Datenstruktur vorliegen. Hierbei wird mit Datenstruktur der Aufbau der Datensätze mit den Attributen gemeint, womit ein Data-Mining Werkzeug ohne Probleme arbeiten kann um das Analyseziel zu erreichen. Die ursprüngliche Datenstruktur umfasst in einem Attribut alle Prozessparameter und die Werte in einem separaten Attribut. Die vertikale Struktur stellt sich schwierig für eine DM-Analyse sowie für andere Auswertungen. Eine horizontale Attributauflistung der Prozessparameter kann von Data-Mining Programmen gut gelesen werden und auch eindeutig zugeordnet werden. Des Weiteren wird nur die PSN des Nutzens übertragen, die restlichen PSN der einzelnen Platinen werden im Reflow-Ofen nicht erfasst. Die Qualitätsbeurteilung beschränkt sich nicht allein auf den Nutzen, sondern auf jede einzelne Platine, daher werden zusätzlich die einzelnen PSN benötigt. Die PSN der Leiterplatinen sollte bereits bei der Erfassung der PSN des Nutzens zusätzlich und automatisch im MES übertragen werden. Dieser Prozess soll zukünftig Analysen vereinfachen. In **Abbildung 47** ist die jetzige MES Datenstruktur und darunter zusätzlich die vorgeschlagene Struktur (s. **Abbildung 48**) abgebildet.

Z1_TEMP_TOP	195	6002292303
Z1_TEMP_BOT	131	6002292303
Z2_TEMP_TOP	165	6002292303
Z2_TEMP_BOT	144	6002292303
Z3_TEMP_TOP	195	6002292303
Z3_TEMP_BOT	168	6002292303
Z4_TEMP_TOP	200	6002292303
Z4_TEMP_BOT	177	6002292303
Z5_TEMP_TOP	260	6002292303
Z5_TEMP_BOT	260	6002292303
Z6_TEMP_TOP	260	6002292303
Z6_TEMP_BOT	260	6002292303
Z7_TEMP_TOP	70	6002292303
TRACK1_TRANSPORT_SPEED	900	6002292303
TRACK1_TRANSPORT_WIDTH	240800	6002292303
TRACK1_PCB_SUPPORT_WIDTH	120000	6002292303
TRACK1_PCB_SUPPORT_HEIGHT	0	6002292303
MACHINE_NITROGEN	962	6002292303
MACHINE_COOLING_BATTERY_TOP	1	6002292303

Abbildung 47: Jetzige Datenstruktur im MES

PSN	Z1_TEMP_TOP	Z1_TEMP_BOT	Z2_TEMP_TOP
6002292303	195	131	165
6002292304	195	131	165
6002292305	195	131	165
6002292306	195	131	165
6002292307	195	131	165
6002292308	195	131	165
6002292309	195	131	165

Abbildung 48: Vorgeschlagene neue Datenstruktur

5.2.2 Datentechnische Prozessvernetzung

Um eine DM-Analyse erfolgreich durchführen zu können, werden Daten benötigt. Die Zugänglichkeit zu Daten von Maschine, Anlage oder In-Line Qualitätskontrollen sollte immer den Berechtigten ermöglicht werden. Die Anlagen und Maschinen sind größtenteils am MES angeschlossen, daher besteht die Möglichkeit von überall auf die Daten zu zugreifen. Jedoch liegt das Problem vor, dass die Qualitätsdaten von In-Line-Kontrollen auf dem lokalen Rechner abgelegt werden. Die Zugänglichkeit wird erschwert und Fachleute müssen anwesend sein. Um dies zu vermeiden, wäre es sinnvoll und wichtig die Daten auf einen zentralen Server oder im MES zu hinterlegen. Der Zugriff auf die Daten wäre weltweit möglich und ist nicht gebunden, am Produktionsstandort anwesend zu sein. Um Analysen effektiv durchzuführen zu können, sollten Daten kompakt z.B. in Excel-Format exportierbar sein.

6. Zusammenfassung und Ausblick

In dieser Arbeit wurde das Data-Mining Verfahren auf Prozessdaten, die von der Wilo SE zur Verfügung gestellt, angewendet. Die Kernaufgabe bestand darin, anhand der DM-Analyse die Qualität der Elektronikplatinen zu optimieren, indem die Soll-Prozessparameter untersucht und eventuell angepasst werden. In Kapitel 2 wurden die notwendigen Grundlagen in den Themenfelder Fertigung, Qualitätsmanagement und Informationssysteme vorgestellt. In Kapitel 3 wurden die erforderlichen Grundlagen von Knowledge Discovery in Database, Data Mining sowie die Methoden der Datenvorverarbeitung und Data-Mining Verfahren erläutert. Die Struktur für die Wissensgewinnung ist in dieser Masterarbeit nach dem Vorgehensmodell MESC aufgebaut, die vom Lehrstuhl ITPL zur Verfügung stellt. In Kapitel 4 wird ein Einblick die Produktionsstruktur der Elektronikfertigung gewährt und die Firmendaten, mithilfe von Datenvorverarbeitungs-methoden, auf das Zielformat gebracht. Anschließend erfolgt die Modellierung von DM-Prozessen für die drei Methoden: Entscheidungsbaum, Clustern und Assoziationsanalyse. Für die Ermittlung der passenden Methoden, wurden drei Methoden auf die Daten angewendet und bewertet. Zum Schluss wurden Handlungsempfehlungen für die weitere Optimierung vorgeschlagen.

Die erste Herausforderung zur Bewältigung der Aufgaben bestand darin, den Untersuchungsbereich einzuschränken, den Prozess zu verstehen und sich die nötigen Daten zu besorgen. Des Weiteren auftretende Fragen mit Experten klären um die Systeme besser zu verstehen.

Die zweite Herausforderung stellte sich bei der ersten Analyse der Reflow-Ofen Datentabelle, alle Prozessparameter sind unter einem Attribut untergeordnet. Die vertikale Ordnung ist für die DM-Analyse schwierig, um die Relation zu Qualitätsausprägung zu erstellen. Jeder Prozessparameter wird ein eigenes Attribut.

Die dritte Herausforderung dieser Arbeit war die Zusammenfassung der Datentabellen vom Reflow-Ofen und AOI. Hierbei bestand die Problematik bei der Erfassung der Product Serial Number. Im Reflow-Ofen wird die PSN vom Nutzen und bei der Qualitätskontrolle, die PSN von jeder einzelnen Platine, erfasst und in der Datenbank hinterlegt. Demzufolge war die Zusammenfassung der Prozessparameter und Qualitätsausprägungen der einzelnen Platinen, aufgrund des fehlenden Schlüsselattributs, nicht möglich. Die Qualitätsausprägung kann nicht jeder einzelnen Platine exakt zugeordnet werden. Deshalb die Ergänzung anhand von Excel nötig.

Die vierte Herausforderung besteht in der Einarbeitung in ein DM-Programm: RapidMiner und die Modellierung der Methoden, um die Analyse zu ermöglichen. Des Weiteren sind Änderungen bezüglich des Datentyps notwendig, die jedoch vom Verfahren abhängig ist.

Die letzte Herausforderung war die Interpretation der Modellierungsergebnisse. Hierbei war neben der Bewertung, auch die Frage zu klären, welche Methode sich für die Analyse der Produktionsdaten gut eignen.

Literatur

[ABEL11]

Abele, Eberhard; Reinhart, Gunther; *Zukunft der Produktion: Herausforderungen, Forschungsfelder, Chancen,* 2011, S.5-37.

[APEL15]

Apel, Detlef; *Datenqualität erfolgreich steuern: Praxislösungen für Business-Intelligence-Projekte,* 3. Auflage, Heidelberg. dpunkt-Verl., 2015, S.1-3.

[BEHN08]

Behnisch, Martin; *Urban Data Mining: Operationalisierung der Strukturerkennung und Strukturbildung von Ähnlichkeitsmustern über die gebaute Umwelt,* Karlsruhe. Universitätsverlag, 2008, S.95-96.

[BERN11]

Bernard, Thomas; *Data Mining im Produktionsumfeld Microsoft PowerPoint - Bernard_IOSB.ppt: Steigerung der Prozess-Effizienz und Produktqualität mit Data-Mining-Methoden,* 2011.

[BODE06]

Bodendorf, Freimut; *Daten- und Wissensmanagement,* Berlin Heidelberg. Springer-Verlag Berlin Heidelberg, 2006, S.1-5.

[BODR03]

Bodrow, Wladimir; Bergmann, Philipp; *Wissensbewertung in Unternehmen: Bilanzieren von intellektuellem Kapital,* Berlin. Schmidt, 2003, S.15-49.

[BOSS04]

Bossel, Hartmut; *Systeme, Dynamik, Simulation: Modellbildung, Analyse und Simulation komplexer Systeme,* Norderstedt. Books on Demand, 2004, S.35-38.

[BRÜG12]

Brüggemann, Holger; Bremer, Peik; *Grundlagen Qualitätsmanagement: Von den Werkzeugen über Methoden zum TQM,* Wiesbaden. Springer Vieweg [+ Teubner], 2012, S.105-109 und S.213.

[BUNG12]

Bungartz, Oliver; *Handbuch Interne Kontrollsysteme (IKS): Steuerung und Überwachung von Unternehmen,* 3. Auflage, Berlin. Schmidt, 2012, S. 125-126.

[CHAM06]

Chamoni, Peter; Beekmann, Frank (Hrsg.); *Analytische Informationssysteme: Business-Intelligence-Technologien und -Anwendungen ; mit 13 Tabellen,* 3. Auflage, Berlin u.a. Springer, 2006, S.61-70.

[CHAP2000]

Chapman, Pete; Clinton, Julian; Kerber, Randy; Khabaza, Thomas; Reinartz, Thomas; Shearer Colin; Wirth, Rüdiger; *CRISP-DM 1.0 - Step-by-step data mining guide.SPSS Inc.,* 2000.

[CLEV16]

Cleve, Jürgen; Lämmel, Uwe; *Data mining,* 2. Auflage, 2016, S.33-228.

[CRON10]

Crone, Sven F.; *Neuronale Netze zur Prognose und Disposition im Handel,* Wiesbaden. Gabler Verlag / GWV Fachverlage GmbH Wiesbaden, 2010,S.189ff.

[DEUS13]

Deuse, Jochen; Erohin, Olga; Lieber, Daniel; *Wissensentdeckung im industriellen Kontext: Herausforderungen und Anwendungsbeispiele.: Wissensentdeckung in vernetzten, industriellen Datenbeständen*, in: Zeitschrift für wirtschaftlichen Fabrikbetrieb, 108(06), S. 388–393.

[DIN95]

DIN EN ISO 8402 Beiblatt 1:1995-08; *Qualitätsmanagement - Anmerkungen zu Begriffen,* 1995.

[DIN15]

DIN EN ISO 9000:2015-11; *Qualitätsmanagementsysteme - Grundlagen und Begriffe,* 2015.

[DREW10]

Drews, Günter; Hillebrand, Norbert; *Lexikon der Projektmanagement-Methoden: [die besten Methoden für jede Situation ; Werkzeugkasten für effizientes Projektmanagement ; auf CD-ROM: Methodenbeispiele und Checklisten],* 2. Auflage, Freiburg u.a. Haufe Mediengruppe, 2010, S.61-75.

[ESTE00]

Ester, Martin; Sander, Jörg; *Knowledge discovery in databases: Techniken und Anwendungen,* Berlin u.a. Springer, 2000,S.109.

[FAYY96]

Fayyad, Usama; Piatetsky-Shapiro Gregory; Smyth Padhraic; *From Data Mining to Knowledge Discovery in Databases*, in: AL Magazine Volume 17, 1996, S. 37–54.

[GAUB09]

Gaubinger, Kurt; Werani, Thomas; Rabl, Michael; *Praxisorientiertes Innovations- und Produktmanagement: Grundlagen und Fallstudien aus B-to-B-Märkten,* 1. Auflage, Wiesbaden. Gabler, 2009, S.130-137.

[GÖTZ13]

Götzfried, Alexander; *Analyse und Vergleich fertigungstechnischer Prozessketten für Flugzeugtriebwerks-Rotoren,* München. Utz, 2013, S.10-29.

[GRÖG15]

Gröger, Christoph; *Advanced Manufacturing Analytics: Datengetriebene Optimierung von Fertigungsprozessen,* 1. Auflage, Lohmar, Rheinl. Eul, J, 2015, S.17-29.

[HACH10]

Hachtel, Günther; Holzbaur, Ulrich D.; *Management für Ingenieure: Technisches Management für Ingenieure in Produktion und Logistik ; mit 73 Tabellen,* 1. Auflage, Wiesbaden. Vieweg + Teubner, 2010, S.67ff.

[HATZ09]

Hatzinger, Reinhold; Nagel, Herbert; *PASW Statistics: Statistische Methoden und Fallbeispiele ; [ehemals SPSS ; www.pearson-studium.de ; companion website],* München. Pearson Studium, 2009, S.31-35.

[HEPP08]

Hepp, Christina; *Fehler- und Fehlerfolgekosten in Banken: Messung und Steuerung der internen Dienstleistungsqualität,* 1. Auflage, Wiesbaden. Gabler, 2008, S.65-110.

[HERI96]

Hering, Ekbert; Triemel, Jürgen; *CAQ im TQM: Rechnergestütztes Qualitätsmanagement,* Wiesbaden. Vieweg+Teubner Verlag; Imprint, 1996,S.112.

[HERR16]

Herrmann, Joachim; Fritz, Holger; *Qualitätsmanagement: Lehrbuch für Studium und Praxis,* 2. Auflage, 2016,S.1-72.

[JUNG13]

Jung, Berndt; Schweißer, Stefan; Wappis, Johann; *Qualitätssicherung im Produktionsprozess,* München. Hanser, 2013, S.21-49.

[KEUP09]

Keuper, Frank; Neumann. Fritz (Hrsg.); *Wissens- und Informationsmanagement: Strategien, Organisation und Prozesse,* 1. Auflage, Wiesbaden. Gabler, 2009, S.151ff.

[KIEH01]

Kiehl, Peter; Klein, Martin; *Klein Einführung in die DIN-Normen,* 13. Auflage, Wiesbaden. Vieweg+Teubner Verlag, 2001, S.315ff.

[KING14]

King, Stefanie; *Big Data: Potential und Barrieren der Nutzung im Unternehmenskontext,* Wiesbaden. Springer Fachmedien Wiesbaden, 2014,S.19-25.

[KLET07]

Kletti, Jürgen; *Konzeption und Einführung von MES-Systemen: Zielorientierte Einführungsstrategie mit Wirtschaftlichkeitsbetrachtungen, Fallbeispielen und Checklisten,* Berlin u.a. Springer, 2007,S.1-12 und S.57.

[KLET15]

Kletti, Jürgen (Hrsg.); *MES - Manufacturing Execution System: Moderne Informationstechnologie unterstützt die Wertschöpfung,* 2. Auflage, Berlin, Heidelberg. Springer Vieweg, 2015, S.31-78.

[KLOS01]

Klosa, Oliver; *Wissensmanagementsysteme in Unternehmen: State-of-the-Art des Einsatzes,* 2001, S,6-36.

[KOUK01]

Koukal, Claus-Ekkehard; *Informationsdarstellung und Wissensverarbeitung in der Arbeitsorganisation prozessnaher Tätigkeiten in Webereien,* Renningen-Malmsheim. Expert, 2001, S.31-45.

[KURB16]

Kurbel, Karl; *Enterprise Resource Planning und Supply Chain Management in der Industrie: Von MRP bis Industrie 4.0,* 8. Auflage, Berlin, Boston. De Gruyter Oldenbourg, 2016, S.40ff.

[OETZ05]

Oetzmann, Arne; *Einsatz wissensbasierter Systeme in Qualitätsmanagement von Produktionsverbünden,* Essen. Vulkan-Verl., 2005, S.7-45.

[PETE05]

Petersohn, Helge; *Data Mining: Verfahren, Prozesse, Anwendungsarchitektur,* München, Wien. Oldenbourg, 2005, S.4-172.

[Pfei14]

Pfeifer, Tilo; Schmitt, Robert (Hrsg.); *Masing Handbuch Qualitätsmanagement,* 6. Auflage, München. Hanser, 2014, S.1-13.

[PROB12]

Probst, Gilbert; Raub, Steffen; Romhardt, Kai; *Wissen managen: Wie Unternehmen ihre wertvollste Ressource optimal nutzen,* 7. Auflage, Wiesbaden. Springer Gabler, 2012, S.1-12.

[Rapi10]

RapidMiner 5.0-Benutzerhandbuch, S. 19–21.

[RUNK10]

Runkler, Thomas A.; *Data Mining: Methoden und Algorithmen intelligenter Datenanalyse,* Wiesbaden. Vieweg + Teubner, 2010, S.23-36.

[SCHE17]

Scheidler, Anne Antonia; *Methode zur Erschließung von Wissen aus Datenmustern in Supply-Chain-Datenbanken,* Göttingen. Cuvillier Verlag, 2017, S.1-123.

[SCHM13]

Schmelzer, Hermann J.; Sesselmann, Wolfgang; *Geschäftsprozessmanagement in der Praxis: Kunden zufriedenstellen, Produktivität steigern, Wert erhöhen,* 8. Auflage, München. Hanser, 2013. S.51-77

[SCHM02]

Schmidt, Günter; *Prozeßmanagement: Modelle und Methoden,* 2. Auflage, Berlin, Heidelberg. Springer Berlin Heidelberg, 2002, S.51-59.

[SCHM15]

Schmitt, Robert; Pfeifer, Tilo; *Qualitätsmanagement: Strategien - Methoden - Techniken,* 5. Auflage, 2015, S.17-36.

[SCHN08]

Schneider, Gabriel; Geiger, Ingrid Katharina; Scheuring, Johannes; *Prozess- und Qualitätsmanagement: Grundlagen der Prozessgestaltung und Qualitätsverbesserung mit zahlreichen Beispielen, Repetitionsfragen und Antworten,* 1. Auflage, Zürich. Compendio Bildungsmedien, 2008, S.174-184.

[SCHÖN16]

Schön, Dietmar; *Planung und Reporting: Grundlagen, Business Intelligence, Mobile BI und Big-Data-Analytics,* 2. Auflage, 2016, S.1-26

[SHAR13]

Sharafi, Armin; *Knowledge Discovery in Databases: Eine Analyse des Änderungsmanagements in der Produktentwicklung,* Wiesbaden. Springer Fachmedien Wiesbaden, 2013, S.48-98.

[VDI13]

VDI; *VDI-Richtlinien 5600 (2013)*.

[WERN04]

Werner, Matthias; *Einflussfaktoren des Wissenstransfers in wissensintensiven Dienstleistungsunternehmen: Eine explorativ-empirische Untersuchung bei Unternehmensberatungen,* 1. Auflage, Wiesbaden. Dt. Univ.-Ver, 2004, S19-23.

[WEST06]

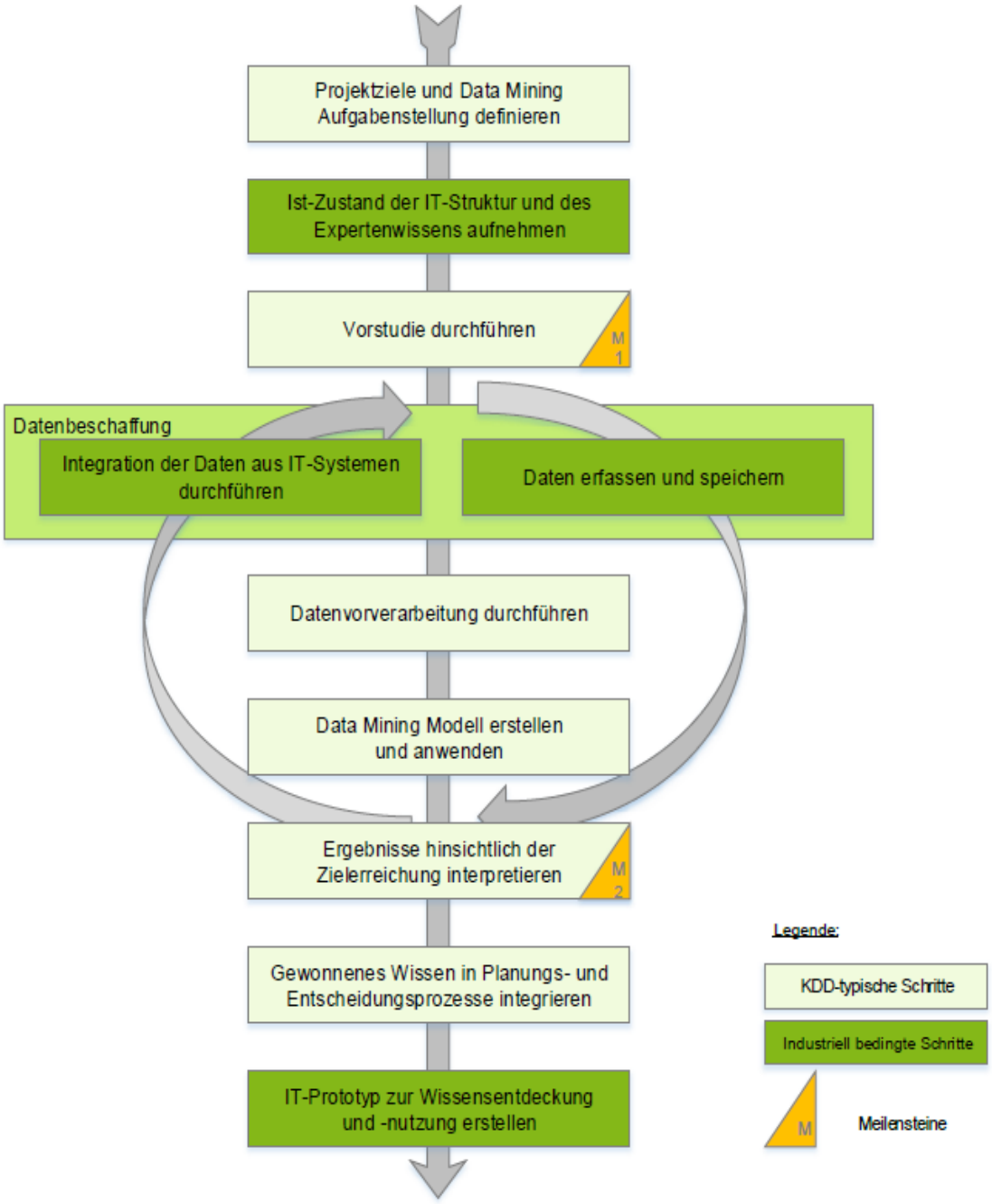
Westkämper, Engelbert; *Einführung in die Organisation der Produktion,* Berlin u.a. Springer, 2006, S.233-242.

[WEST16]

Westkämper, Engelbert; Löffler, Carina; *Strategien der Produktion: Technologien, Konzepte und Wege in die Praxis,* Berlin, Heidelberg. Springer Vieweg, 2016, S.45-108 und S.239ff.

Anhang

Anhang 1: Knowledge Discovery in Industrial Database [DEUS13]



Anhang 2: Für die Datenvorverarbeitung benutzte Makros

Makro für das Einfügen von Leerzeilen, für die Ergänzung der PSN der einzelnen Platinen, in Excel

```
Sub Leerzeilen()
Dim i As Integer
Dim z As Integer
Application.ScreenUpdating = False
For i = Cells(Rows.Count, 1).End(xlUp).Row To 1 Step -1
For z = 1 To 6
Cells(i, 1).EntireRow.Insert Shift:=xlDown
Next z
Next i
Application.ScreenUpdating = True
End Sub
```

Makro für die Ergänzung der PSN der einzelnen Platinen in Excel

```
Sub RPP()
Dim a As Range
With Tabelle1
With Range(.Cells(1, 2), .Cells(.Rows.Count,1).End(xlUp)).SpecialCells(xlCellTypeBlanks)
   .FormulaR1C1 = "=R[-1]C1+1"
   For Each a In .Areas
     a.Copy: a.PasteSpecial xlPasteValues
   Next
 End With
End With
Application.CutCopyMode = False
End Sub
```

Anhang 3: SMD-Reflowofen- Sollprozessdaten

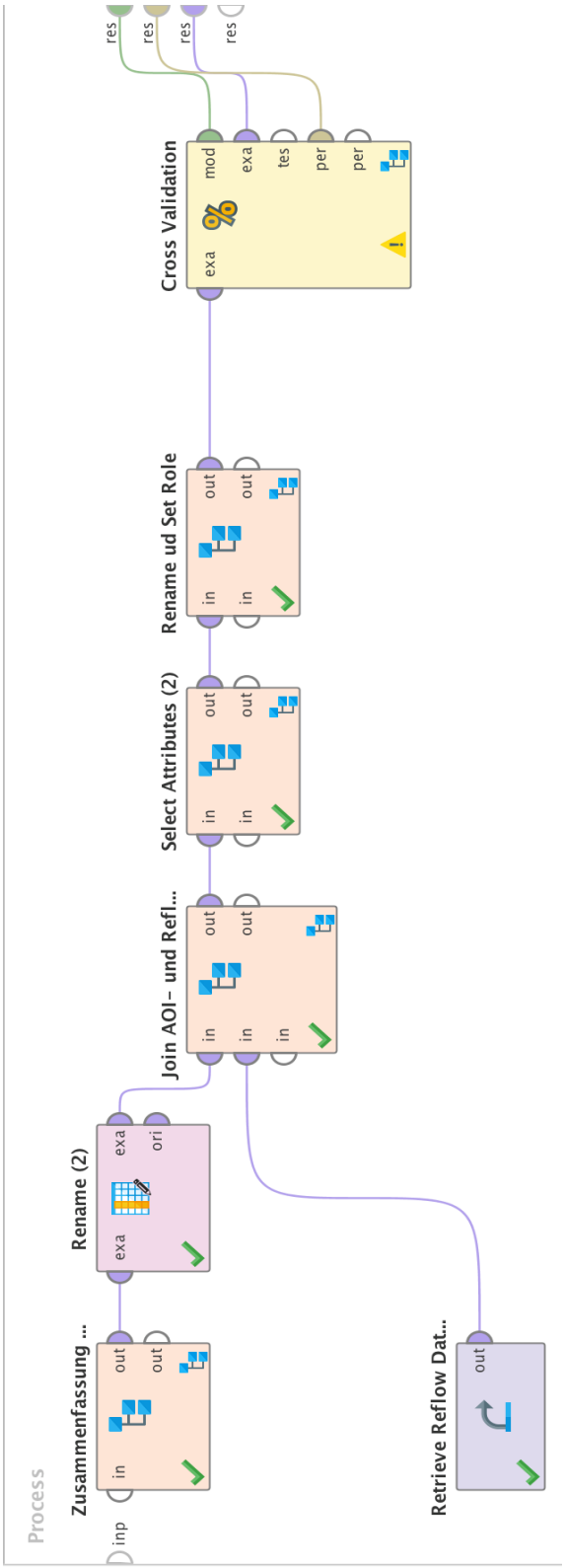
Sollprozessdaten vom Unternehmen in der Anlage eingestellt

	Zone 1	Zone 2	Zone 3	Zone 4	Peakzone1	Peakzone 2	Kühlzone
Top	195	165	180	200	260	260	
							80
Bot	195	165	180	200	260	260	

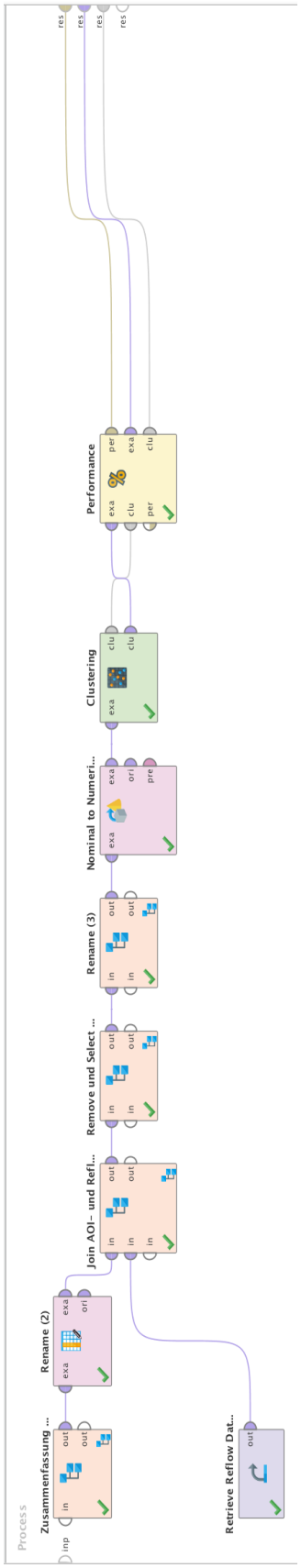
O2 Peak obere Maximalwert 2000 ppm

1000 ppm---- Sollwert untere Minimalwert 500 ppm

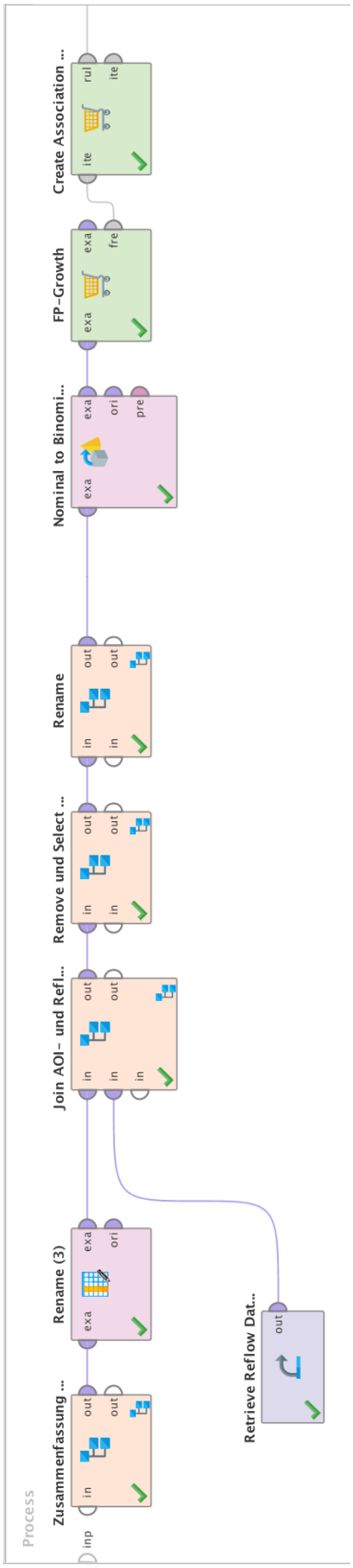
Anhang 4: Kompletter Prozess mit dem ID3 Algorithmus



Anhang 5: Kompletter Prozess mit dem k-Means Algorithmus



Anhang 6: Kompletter Prozess mit dem FP-Growth Algorithmus



Eidesstattliche Versicherung