HOCHSCHULE HEILBRONN

Wissenschaftlicher Beitrag

Metaleaning-Ansatz zur Entwicklung eines Vorgehensmodells im Rahmen von CRISP-DM zum Einsatz von KI-Methoden mit Fallbeispielen aus Stammdatenmanagement

Victor Pontello[1]

[1] Hochschule Heilbronn, Studiengang Wirtschaftsinformatik – Informationsmanagement und Data Science, vpontell@stud.hs-heilbronn.de

19.07.2020 – Version 1

Abstract

Die Umsetzung von Künstliche Intelligenzmethoden (KI-Methoden) ist oft keine triviale Aufgabe. Besonders bei Anfänger ist es eine große Herausforderung, eine passende KI-Methode zu einer bestimmten Problemstellung herauszufinden (Zschech *u. a.*, 2020, S. 228). Darüber hinaus ist ein heuristischer Ansatz zur Auswahl von Algorithmen oft fehleranfällig und suboptimal (Ali, Lee und Chung, 2017, S. 257). Als Lösung für dieses Problem wurde auf Basis eines Metalearning-Ansatzes ein Vorgehensmodell im Rahmen von CRISP-DM entwickelt. Bei dem Modell wird anhand eines Kriterienrasters Problemstellung charakterisiert und auf diesen Charakteristiken aufbauend ist das Vorgehensmodell in der Lage, optimale KI-Methoden als Lösungsmöglichkeiten vorzuschlagen. Das Modell wurde anhand von Aufgaben aus Stammdatenmanagement evaluiert und die beobachteten Ergebnisse finden Begründung und Berücksichtigung in der Literatur.

Keywords: *KI, Machine Learning, Stammdatenmanagement, Metalearning, CRISP-DM*

1 Einleitung

1.1 Problemstellung

Die zunehmende Datenmenge stellt die Datenverwaltung von Unternehmen vor Herausforderungen (Otto, Hüner und Österle, 2012). Obwohl die Firmen die KI als eine interessante Gelegenheit für Mehrwertgenerierung halten, fehlt oft nicht nur das Fachwissen zur Auswahl einer passenden KI-Methode, aber auch bezüglich ihrer Anwendungsmöglichkeiten (Brazdil *u. a.*, 2009, S. 2). Unter KI-Methoden wird ein breiteres Spektrum von Methoden verstanden, die für unterschiedliche Anwendungen geeignet sind und unterschiedliche Voraussetzungen enthalten (Ali, Lee und Chung, 2017, S. 257). Gleichzeitig wird es, insbesondere für Neulinge, immer komplexer, den Überblick zu behalten und zu entscheiden, welche KI-Methoden in welchem Kontext am besten geeignet sind (Zschech *u. a.*, 2020, S. 228). Oft wird diese Aufgabe von gut ausgebildeten qualifizierten KI-Experten gelöst, die jedoch eine seltene und kostenintensive Belegschaft sind (Zschech *u. a.*, 2018, 2020, S. 228). Darüber hinaus kann eine falsche Auswahl der KI-Methode zu Misserfolgen und exzessiven Kosten- und Zeitaufwand führen (Ali, Lee und Chung, 2017, S. 257).

1.2 Fragestellung

Auf diesem Kontext aufbauend kann die folgende Forschungsfrage abgeleitet werden:

> Wie kann die Auswahl von KI-Methoden auf Basis der Charakteristiken des zu lösenden Problems effizient durchgeführt werden?

1.3 Zielsetzung

In dieser Arbeit wird als Artefakt ein Vorgehensmodell im Rahmen des CRISP-DM als Beantwortung der obengenannten Forschungsfragen entwickelt. Eine standardisierte Vorlage für die Problemcharakterisierung wird in der Form eines Kriterienrasters geliefert, sodass die wesentlichen Informationen über das Problem als Eingabe zum Vorgehensmodell verwendet werden. Diese Informationen werden anhand eines Entscheidungsbaums bearbeitet, sodass für eine bestimmte Problemstellung optimale KI-Methoden als Lösungsvorschlag ausgegeben werden. Nachher wird das Vorgehensmodell mit drei typischen Problemen aus Stammdatenmanagement getestet und die Lösungsvorschläge werden ausgewertet.

1.4 Aufbau der Arbeit

Die Forschungsstudie wird im Rahmen des „Design Science Reseach"-Modells von Peffers (2007) aus einem problem- und zielorientierte Initiative durchgeführt (Peffers *u. a.*, 2007). Die Problembeschreibung und Motivation sind in der Einleitung (S. 2), sowie die Zielsetzung und Lösungsansatz beschrieben. Diese werden zusätzlich im Abschnitt Grundlagen und Begriffe (S. 3) auf Basis von der Literatur und Theorie diskutiert und vertieft. Sowohl das Design und Entwicklung des Artefaktes zur Beantwortung der Forschungsfrage als auch eine Anwendungsdemonstration werden im Abschnitt Ergebnisse (S. 9) umgesetzt. Anschließend wird im Abschnitt Diskussion und Ausblick (S. 18) eine kritische Evaluation mit Blick auf zukünftige Forschungen ausgeführt.

2 Grundlagen und Begriffe

In diesem Abschnitt wird die Wissensbasis dieser Arbeit dargestellt. Das Kapitel wurde in drei wesentlichen Themengebiete aufgeteilt. Erstens das Rahmenwerk, auf Basis dessen das Vorgehensmodell aufgebaut ist (CRISP-DM). Zweitens die KI-Methoden und ihre Grundlagen, die der Ausgangspunkt und Kern des Vorgehensmodells sind. Anschließend das Themengebiet Stammdatenmanagement, aus dem Problemstellungen zur Evaluation des Vorgehensmodells stammen.

2.1 CRISP-DM

Der Einstieg und Durchführung von datenbasierten Projekten kann durch die Anwendung eines standardisierten und strukturierten Vorgehensmodell erleichtert werden, indem die Planung und Verwaltung des Projektes dadurch unterstützt werden und somit seine Validität und Zuverlässigkeit erhöht werden In diesem Kontext hebt das Modell CRISP-DM (*Cross Industry Standard Process for Data Mining*) hervor, als ein sehr weit verbreitetes und ausgereiftes Prozessmodell zur Durchführung von Datenanalysen (Schacht und Lanquillon, 2019, S. 112).

Das Modell CRISP-DM bietet jedem, vom Anfänger bis zum Data-Mining-Experten, ein vollständiges Layout für den Entwurf eines Data-Mining-Projekts (Shearer *u. a.*, 2000, S. 14). Das vollständige Modell ist in der Abbildung 1 mit seinen sechs Phasen dargestellt. Allerdings liegt der Fokus im Rahmen dieser Arbeit auf den ersten vier Phasen: *Business Understanding. Data Understanding, Data Preparation* und *Data Modeling,* die folgend näher erläutert werden.

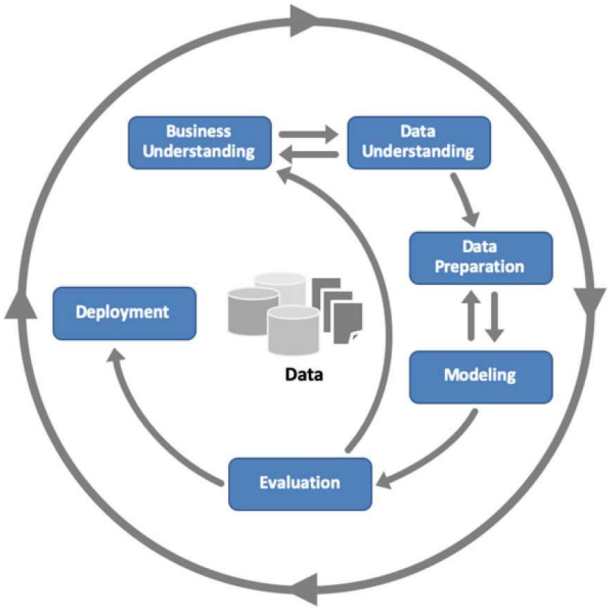

Abbildung 1: Darstellung des CRISP-DM-Modells und seine sechs Phasen (Schacht und Lanquillon, 2019, S. 112)

Die erste Phase von CRISP-DM ist das Geschäftsverständnis oder *Business Understanding.* Der Fokus in dieser Phase liegt darauf, die Projektziele aus einer Geschäftsperspektive zu verstehen und eine Data-Mining-Problemdefinition daraus abzuleiten. Darauf aufbauend wird einen vorläufigen Plan zur Erreichung der Ziele entwickelt. Zu den wichtigsten Schritten dieser Phase gehören die Festlegung der Geschäftsziele, die Bewertung der Situation, die Festlegung der Data-Mining-Ziele und die Erstellung des Projektplans (Shearer *u. a.*, 2000, S. 14).

Bei der Phase *Data Understanding* oder Datenverständnisses wird erst eine Datensammlung umgesetzt. Hervorhebende Merkmale sowie Datenqualitätsprobleme werden identifiziert und Hypothesen vorhandene oder verborgene Informationen werden gebildet. Bei dieser Phase werden vier Schritte durchgeführt, darunter die Sammlung der Ausgangsdaten, die Beschreibung der Daten, die Untersuchung der Daten und die Überprüfung der Datenqualität (Shearer *u. a.*, 2000, S. 15).

Die Datenvorbereitungsphase oder *Data Preparation* umfasst alle Aktivitäten zur Erstellung des endgültigen Datensatzes oder der Daten, die aus den anfänglichen Rohdaten in das Modellierungswerkzeug eingespeist werden. Die fünf Schritte der Datenaufbereitung sind die Auswahl der Daten, die Bereinigung der Daten, die Konstruktion der Daten, die Integration der Daten und die Formatierung der Daten (Shearer *u. a.*, 2000, S. 16).

Bei der Modellierungsphase oder *Data Modeling* werden verschiedene Modellierungs- und Optimierungstechniken ausgewählt. Einige dieser Techniken haben spezifische Anforderungen an die Form der Daten, was zu einem Rückschritt in den vorherigen Phasen führen kann. Zu den Modellierungsschritten gehören die Auswahl der Methode, die Generierung des Testdesigns, die Erstellung von Modellen und die Bewertung der Modelle (Shearer *u. a.*, 2000, S. 17).

Was in dieser Arbeit im Rahmen von CRISP-DM abgezielt ist, ist eine Verbindung zwischen den Phasen aus *Business Understanding* über die *Data Understanding* und die *Data Pretaration* hin zu *Data Modeling* mittels eines Vorgehensmodells in Data Mining Projekten zu Realisieren. Der Schlüsselaspekt dieses Projekt ist, das Potenzial der Einsetzung von KI-Methoden als Problemlösung in datengetriebenen Problemen zu entdecken. Aus dieser Perspektive sind einige Schritte aus diesem Prozessmodell besonders relevant. Z.B. die Festlegung der Data-Mining-Ziele und Erstellung des Projektplans (*Business Undestanding*), die Beschreibung der Daten und Überprüfung der Datenqualität (*Data Understanding*) und Integration der Daten (*Data Preparation*). Aus den Informationen bezüglich dieser Schritte kann die Auswahl der Methode (*Data Modeling*) anhand eines Vorgehensmodells schon vorab durchgeführt werden sodass die optimalen Lösungsmöglichkeiten und -Potenzial der KI-Methoden bewusst sind.

2.2 Metalearning

Metalearning ist ein Themengebiet aus Machine Learning, bei dem abgezielt wird, die Benutzer bei der Aufgabe der Auswahl von KI-Methoden unter Berücksichtigung des Anwendungsbereichs zu unterstützen (Brazdil *u. a.*, 2009, S. 1). Auf Basis dieses Themengebiet wird das Vorgehensmodell entwickelt.

2.3 KI-Methoden

Was in dieser Arbeit als KI-Methode bezeichnet wird, könnte auch als Machine Learning Verfahren, oder als Lernverfahren bezeichnet werden. Jedoch wird der Begriff KI-Methoden zwecks Einheitlichkeit am meisten in dieser Arbeit verwendet. In diesem Sinne werden in diesem Abschnitt unterschiedlichen KI-Methoden charakterisiert und unterschieden.

2.3.1 Definition

Die KI-Methoden lassen sich durch die Definition von Maschinelles Lernen gut definieren, und zwar nach Arthur L. Samuel (Samuel, 1959) ist Maschinelles Lernen das Studiengebiet, das Computern die Fähigkeit verleiht, ohne explizite Programmierung zu lernen. Die sogenannte

Wissenschaft des Lernens in der Hinsicht der KI-Methoden spielt eine Schlüsselrolle in den Bereichen Statistik, Data Mining und künstliche Intelligenz und überschneidet sich mit Bereichen des Ingenieurwesens und anderen Disziplinen (Hastie, Tibshirani und Friedman, 2009, S. 1).

2.3.2 *No Free Lunch*

Jedoch gibt es in Bezug auf KI-Methoden kein Lernverfahren, das „das Beste“ ist. Die Anwendung eines Lernverfahrens ist fallabhängig und je nachdem wie die Rahmenbedingungen der abgezielte Problemlösung sind, können unterschiedliche Lernverfahren ausgewählt werden (Wolpert, 1996, S. 1414). Es wurde von wissenschaftlichen Beiträgen nachgewissen, dass unterschiedliche KI-Methoden weitere KI-Methoden überbieten oder werden von denen überboten, je nach bspw. Eigenschaften und Merkmale der Domänendaten auf den Datensätzen (Caruana und Niculescu-Mizil, 2006; Ali, Lee und Chung, 2017, S. 258).

2.3.3 Struktur zur Charakterisierung von KI-Methoden

Eine definierte Struktur zur Charakterisierung oder Zuordnung von KI-Methoden lässt sich nicht in einer trivialen Form durchsetzen. Unterschiedliche Dimensionen werden dafür von verschiedenen Autoren in der Literatur verwendet (Hastie, Tibshirani und Friedman, 2009; Braschler, Staderlmaan und Stockinger, 2019; Schacht und Lanquillon, 2019; Joshi, 2020; Kreutzer und Sirrenberg, 2020). In dieser Hinsicht werden bspw. nach einigen Literaturquellen die Dimensionen, Lernformen und Aufgabetypen als Ausgangspunkt verwendet, um die Lernverfahren zu definieren (Schacht und Lanquillon, 2019), nach anderen Quellen liegt der Fokus auf der technischen Umsetzung der Algorithmen (Hastie, Tibshirani und Friedman, 2009).

Dadurch, dass keine einheitliche Definition in der Literatur gefunden wurde, wurde für diese Arbeit ein Ensemble aus Charakterisierungskriterien von mehreren Autoren realisiert und darauf aufbauend eine Struktur für die Charakterisierung der KI-Methoden wurde erstellt, wie in der Abbildung 2 dargestellt ist. Diese Struktur wurde für die Entwicklung des Vorgehensmodells später verwendet, wie in den folgenden Abschnitten erklärt ist.

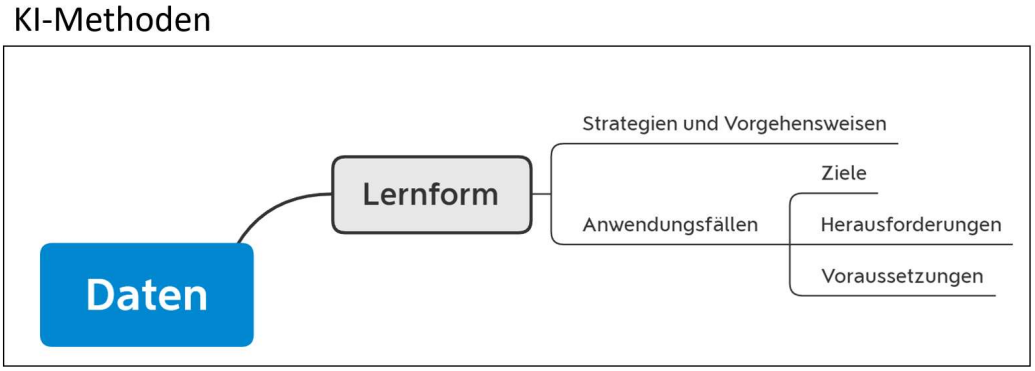

Abbildung 2: Wesentliche Dimensionen für die Charakterisierung von KI-Methoden, (Eigene Darstellung in Anlehnung von (Hastie, Tibshirani und Friedman, 2009; Braschler, Staderlmaan und Stockinger, 2019; Schacht und Lanquillon, 2019; Joshi, 2020; Kreutzer und Sirrenberg, 2020))

2.3.4 Dimensionen der Anwendung einer KI-Methode und ihre Beispiele

In dieser Arbeit werden die Anwendungen von KI-Methoden in vier Dimensionen charakterisiert:

1. **Daten**:

Unter Daten wird verstanden, ob sie strukturiert oder unstrukturiert ist. Als strukturierte Daten werden bspw. Daten in der Form einer Tabelle oder Matrix beschrieben. Unstrukturiert sind die Daten in der Form von natürlichen Sprachen wie Texte und Tone, oder auch Bilder und Videos (Zschech *u. a.*, 2020, S. 229).

2. Lerntyp (oder Lernform):

In der Literatur sind vier Lernformen zu finden. Erstens das Überwachtes Lernen. Sie ist die meistverbreitete Lernform und benötigt eine Zielgröße für das Lernverfahren. Zweitens das Unüberwachtes Lernen. Sie benötigt im Gegensatz zu dem Überwachtes Lernen keine Zielgröße. Nächstens das Halbüberwachtes Lernen als eine Mischform der davor erwähnten Lernformen, bei dem die Zielgrößen nur teilweise zur Verfügung sind. Anschließend das Bestärkendes Lernen, bei dem das Lernen nicht aus den Daten, sondern aus Interaktionen stattfindet (Schacht und Lanquillon, 2019; Joshi, 2020)

3. Methode zum Ziel

Als Methoden zum Ziel werden die Aufgaben bezeichnet, die durch die Anwendung von KI-Methoden gelöst werden können. Wenn kategorische Zielgrößen im Daten vorhanden sind kann eine Klassifikation durchgeführt werden, jedoch wird die Aufgabe als Regression bezeichnet, wenn die Zielgrößen numerisch sind. Falls die Zielgrößen nicht vorhanden sind, sind die wesentlichen Aufgaben das Erkennen von Mustern aus den Daten und Clustering und die Reduzierung der Dimensionen Zwecks Vereinfachung der Daten. Noch dazu kommt der Fall, bei dem das Lernen durch die Interaktionen mit der Umgebung stattfinden. Dafür wird durch Versuch und Irrtum eine Folge von Entscheidungen durch Belohnung und Bestrafung abgebildet auf darauf aufbauend findet das Lernen statt (Schacht und Lanquillon, 2019; Joshi, 2020).

4. Methoden

Als Methoden werden als die Algorithmen verstanden, aus denen die KI-Methoden bestehen und für das Lernen verwendet werden. In dieser Arbeit liegt der Fokus auf die Methodenfamilien. Unter diesen Familien werden verschiedene Algorithmen gruppiert, die auf Basis demselben Paradigma basieren, aber in unterschiedlichen Domäne angewendet werden können (Braschler, Staderlmaan und Stockinger, 2019, S. 107). Somit werden die *Tree-Based-Methods,* wie z.B. Entscheidungsbäume oder Random-Forest, Linear Models, Neuronale Netze und Deep Learning, Bayesian Methode und K-Nearest Neighbors für das Überwachtes oder Halbüberwachtes Lernen in dieser Arbeit abgedeckt. Darüber hinaus werden die K-Means, Hierarchical Clustering, Autoencoding Neuronale Netze und PCA (*Principal Component Analysis*) für das Unüberwachtes Lernen umfasst (Schacht und Lanquillon, 2019). Anschließend sind die Methoden Q-Learning und SARSA für das Bestärkendes Lernen auch im Spektrum dieser Arbeit abgedeckt (Nandy und Biswas, 2018). Eine ausführlichere Umfassung hinsichtlich der technischen Merkmale jeder Methode bzw. Methodenfamilien wird in dieser Arbeit nicht durchgeführt, da der Fokus der Zielsetzung auf der Anwendungsebene liegt und nicht auf den technischen Kennzeichen jeder Methode.

Alle obengenannten Dimensionen und ihre Beispiele sind in der

Daten	Lerntyp	Methode zum Ziel	Methoden
Strukturiert	Überwachtes Lernen	Klassifikation	Tree-Based Methods
Unstrukturiert	Unüberwachtes Lernen	Regression	Linear Models
	Halbüberwachtes Lernen	Vereinfachen durch Dimension Reduction	Support Vector Machines
	Bestärkendes Lernen	Muster erkennen und Clustering	Neural Networks and Deep Learning
		Interaktion statt bereitgestellten Daten	Bayesian Methods
			K-Nearest Neighbors
			K-Means
			Hierarchical Clustering
			Autoencoding NN
			PCA
			SARSA
			Q - Learning

Abbildung 3: Zusammenfassung der Anwendungsdimensionen einer KI-Methode und ihre Beispiele, die in dieser Arbeit abgedeckt sind (Eigene Darstellung) in einer tabellarischen Form dargestellt.

Daten	Lerntyp	Methode zum Ziel	Methoden
Strukturiert	Überwachtes Lernen	Klassifikation	Tree-Based Methods
Unstrukturiert	Unüberwachtes Lernen	Regression	Linear Models
	Halbüberwachtes Lernen	Vereinfachen durch Dimension Reduction	Support Vector Machines
	Bestärkendes Lernen	Muster erkennen und Clustering	Neural Networks and Deep Learning
		Interaktion statt bereitgestellten Daten	Bayesian Methods
			K-Nearest Neighbors
			K-Means
			Hierarchical Clustering
			Autoencoding NN
			PCA
			SARSA
			Q - Learning

Abbildung 3: Zusammenfassung der Anwendungsdimensionen einer KI-Methode und ihre Beispiele, die in dieser Arbeit abgedeckt sind (Eigene Darstellung)

2.4 Stammdatenmanagement

Aus Stammdatenmanagement werden Aufgaben bzw. Probleme abgeleitet, deren Lösungen von dem Vorgehensmodell unterstützt werden. Diese Problemlösungen werden für die Evaluation des Vorgehensmodell in der Form von simulierten Praxisbeispiele angewendet (S.15).

2.4.1 Definition

Das Themengebiet Stammdatenmanagement beschreibt die Daten aus den Kerneinheiten eines Geschäfts, wie bspw. Lieferanten, Kunden, Produkte, Mitarbeiter und Vermögenswerte. Auf diese Daten basieren die Geschäftsaktivitäten eines Unternehmens (Otto, Hüner und Österle, 2012, S. 396 ff.). Darüber hinaus wird Stammdatenmanagement als die Technologie, die Werkzeuge und die Prozess beschrieben, die erforderlich sind, um saubere, konsistente und genaue Listen von Stammdaten zu erstellen und zu pflegen (Das und Mishra, 2011, S. 131).

2.4.2 Konzepte

Die Stammdaten werden in drei Konzeptniveaus zugeordnet: Klassen, Objekte und Attribute. Die Attribute werden durch Werte definiert, wie z.B. die Telefonnummer. Die Stammdatenobjekte stellen ein konkretes Businessobjekts dar, wie bspw. eine Kunde. Die Klassen beschreiben in der Praxis bspw. die Kunden eines Unternehmens. Zu Stammdatenmanagement gehören alle Aufgaben, die die Erstellung, das Löschen und Veränderung einer Stammdatenklasse, -objekt oder -attribute umfasst (Otto, Hüner und Österle, 2012, S. 398 f.).

2.4.3 Funktionsarchitektur aus Stammdatenmanagement

Auf Basis einer funktionalen Ansatz wurde eine Funktionsarchitektur der Stammdatenmanagement entwickelt (Otto und Hüner, 2009), wie in der Abbildung 4 dargestellt ist. In dieser Struktur sind sechs Funktionsgruppen zu finden, unter denen die jeweiligen zugehörigen Funktionen bzw. Aufgaben zugeordnet sind. Aus dieser Struktur können Funktionen oder Aufgaben erkannt werden, die sich sehr an Schritten von CRISP-DM nähen, wie *Data Modeling* (Datenmodellierung), *Data Analysis* (Datenanalyse), *Data Enrichment* (Datenanreicherung) und *Data Cleaning* (Datenbereinigung) (Otto, Hüner und Österle, 2012, S. 418–422).

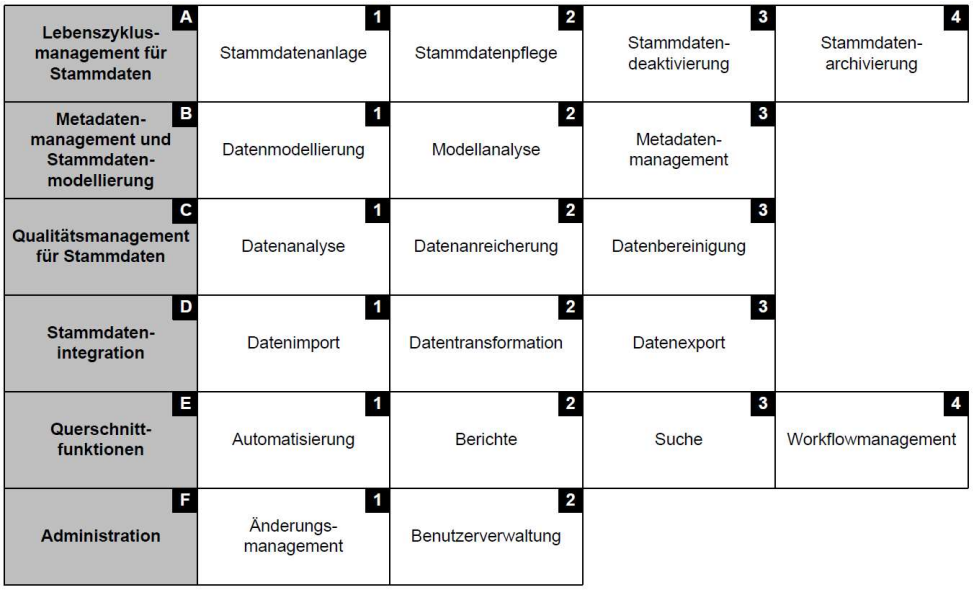

Abbildung 4: Funktionsarchitektur aus Stammdatenmanagement (Otto und Hüner, 2009, S. 21)

Aus den Definitionen der Aufgaben dieser Funktionsarchitektur (Otto und Hüner, 2009, S. 418–422) werden Probleme abgebildet, die mittels des Vorgehensmodells gelöst werden. Diese Anwendungssimulation des Vorgehensmodells gilt als Teil der Evaluationsverfahren und wird im Abschnitt 4.3 (S. 15) näher erläutert.

3 Methodik zur Literaturanalyse

Für die Literaturanalyse wurde die Methodik erstens nach Fettke (Fettke, 2006) und Webster und Watson (Webster und Watson, 2002) eingerichtet. Aufbauend auf diese Methodik wurde eine systematische Literaturanalyse durchgeführt. Bei der die Herausforderung war, der Stand der Technik und die Grundlagen in Bezug auf die KI-Methoden, Stammdatenmanagement und CRISP-DM abzudecken. Als erster Ansatz wurde nach Fettke (Fettke, 2006) vorgegangen, um eine Literaturbasis zu geschaffen und darauf aufbauend wurde diese Basis iterativ nach Webster und Watson durch die Vorwärts- und Rückwärtsrecherche ergänzt (Webster und Watson, 2002).

4 Ergebnisse

Sowohl das Design und Entwicklung des Artefaktes zur Beantwortung der Forschungsfrage als auch eine Anwendungsdemonstration werden in diesem Abschnitt abgedeckt. Das Kapitel ist in

drei wesentliche Schritte aufgeteilt, nämlich die Beschaffung einer Vorlage zur Assoziation zwischen Problemen und KI-Methoden in der Form eines Kriterienrasters (S. 10), die Entwicklung des Vorgehensmodells (S. 12) und die Evaluation des Modells anhand eines Fallbeispiels mit drei Problemlösungen aus Stammdatenmanagement (S. 15).

4.1 Entwicklung der Vorlage zur Assoziation zwischen Problemen und KI-Methoden

4.1.1 Kriterienraster für die Charakterisierung der KI-Methoden

Auf Basis der Literaturrecherche wurden 10 Unterscheidungsmerkmale gefunden, die einen Anwendungsfall einer KI-Methode gut abbilden. In diesem Sinne wurden bei jedem Unterscheidungsmerkmal qualitative Kategorien zugeordnet, um die Methode basierend auf dem Merkmal auszuwerten. Die Entscheidung für eine qualitative Auswertung ist aus dem Grund getroffen, um die Anwendung des Vorgehensmodells zu vereinfachen. weil die quantitativen Werte bspw. der Verfügbarkeit von Zeit nicht immer bewusst und schwer einzuschätzen sind. Die Unterscheidungsmerkmale sind in der Abbildung 5 zusammen mit den qualitativen Klassen zur Charakterisierung der KI-Methoden dargestellt.

Datenformat	Datenqualität	Verfügbarkeit von Zeit	Anspruch auf Genauigkeit	Datentyp	Typ der Zielgröße	Rechenkapazität	Datenmenge	Folge von Entscheidungen	Anzahl an Dimensionen (Features)
Matrix	Sehr Hoch	Sehr Hoch	Sehr Hoch	Gelabelt	Numerisch	Sehr Hoch	Sehr Groß	ja	Gering
Text	Hoch	Hoch	Hoch	Nicht gelabelt	Kategorisch	Hoch	Groß	nein	Mittel
Bild	Normal	Normal	Normal	Gemischt	Keine	Normal	Neutral		Hoch
Video	Gering	Gering	Gering	Feedback-Signal		Gering	Klein		
Ton	Keine						Keine		
Keine									

Abbildung 5: Unterscheidungsmerkmale für die Charakterisierung der KI-Methoden

Die Entscheidung für jedes dieses Merkmales wurde auf Basis der Literatur getroffen. Allerdings müsste für die Umsetzung dieses Projekts eine gewisse Abstraktion geschaffen werden, sodass die hohe Komplexität sowohl die Entwicklung als auch die Nutzung des Vorgehensmodells nicht verhindert. In dieser Art und Weise ist in der Abbildung 6 eine Auflistung der wesentlichen Unterscheidungsmerkmale, ihren Anwendungsgründen und eine Legende mit den Kennzeichen dargestellt.

Die Zusammenbindung von den Unterscheidungsmerkmalen mit den jeweiligen Auswertungskategorien bilden das Kriterienraster ab. In dieser Hinsicht wird dieses Kriterienraster als eine standardisierte Vorlage verwendet, um das zu lösenden Problem zu beschreiben und charakterisierten. Somit ist die Basis für die Assoziation zwischen Problem und Lösungsmethode geschaffen.

Kennzeichen	Unterscheidungsmerkmal	Grund der Anwendung
UM1	Datenformat	Einige KI-Methoden können besser mit unterschiedlichen Datenformate Umgehen als andere.
UM2	Datenqualität	Die Empfindlichkeit auf die Datenqualität kann je nach KI-Methode variieren.
UM3	Verfügbarkeit von Zeit	Der Modellierungsprozess kann je nach KI-Methode mehr oder weniger Aufwand darstellen.
UM4	Anspruch auf Genauigkeit	Einige KI-Methoden können in der Regel weniger Genauigkeit aufweisen, aber gegensätzlich sind sie einfacher zu implementieren.
UM5	Datentyp	Auf Basis dessen wird die Lerntyp bestimmt.
UM6	Typ der Zielgröße	Wichtiges Merkmal bei dem Überwachten Lernen für die Definition der Methode zum Ziel
UM7	Rechenkapazität	Einige KI-Methoden haben mehr Anspruch auf Rechenleistung als andere
UM8	Datenmenge	Bei einigen KI-Methoden ist die Performance mehr oder weniger Abhängig von der Datenmenge
UM9	Folge von Entscheidungen	Wichtig bei der Bestimmung des Lerntyps Bestärkenden Lernen
UM10	Anzahl an Dimensionen (Features)	Einige KI-Methoden oder Methoden zum Ziel sind besser oder schlechter geeignet für das Vorgehen verschiedenen Dimensionen Anzahl

Abbildung 6: Anwendungsgründe der Unterscheidungsmerkmale zur Charakterisierung der KI-Methoden (Eigene Darstellung)

4.1.2 Charakterisierung der einzelnen Methoden und Entwicklung des Basisdatensatzes

Aufbauend auf die im Abschnitt 2.3.4 (S. 5, Abbildung 2) dargestellten Anwendungsdimensionen wurden 35 Anwendungsmöglichkeiten abgebildet wie in der Abbildung 7 zu sehen ist. Aus der Analyse wurde basierend auf der Literaturrecherche in der Form eines *Reverse Engineering* - Ansatzes das Kriterienrasters so beantwortet, dass das Ergebnis die Vorgegebene Anwendungscharakterisierung sein sollte. Dadurch, dass es mehrere mögliche Antworte des Kriterienrasters für eine bestimmte Anwendungscharakterisierung gibt, wurden Kombinationen aus Antworten erstellt, wie in der Abbildung 8 dargestellt ist. Dieses Verfahren wurde bei jeder einzelnen 35 Anwendungsmöglichkeit durchgeführt und am Ende sind aus der Kombinationen 2470 Anwendungsfällen entstanden.

Kennzwichen	Anwendungscharakterisierung einer KI-Methode
AC1	['Strukturiert', 'K-Means', 'Muster erkennen und Clustering', 'Unüberwachtes Lernen']
AC2	['Strukturiert', 'PCA', 'Vereinfachen durch Dimension Reduction', 'Unüberwachtes Lernen']
AC3	['Unstrukturiert', 'Neural Networks and Deep Learning', 'Regression', 'Halbüberwachtes Lernen']
AC4	['Unstrukturiert', 'Q - Learning', 'Interaktion statt bereitgestellten Daten', 'Bestärkendes Lernen']
AC5	['Unstrukturiert', 'SARSA', 'Interaktion statt bereitgestellten Daten', 'Bestärkendes Lernen']
AC6	['Strukturiert', 'Tree-Based Methods', 'Regression', 'Überwachtes Lernen']
AC7	['Strukturiert', 'Neural Networks and Deep Learning', 'Klassifikation', 'Halbüberwachtes Lernen']
AC8	['Strukturiert', 'Autoencoding NN', 'Vereinfachen durch Dimension Reduction', 'Unüberwachtes Lernen']
AC9	['Strukturiert', 'Linear Models', 'Regression', 'Halbüberwachtes Lernen']
AC10	['Strukturiert', 'Neural Networks and Deep Learning', 'Regression', 'Halbüberwachtes Lernen']
AC11	['Strukturiert', 'K-Nearest Neighbors ', 'Regression', 'Halbüberwachtes Lernen']
AC12	['Strukturiert', 'Tree-Based Methods', 'Regression', 'Halbüberwachtes Lernen']
AC13	['Unstrukturiert', 'Neural Networks and Deep Learning', 'Klassifikation', 'Halbüberwachtes Lernen']
AC14	['Strukturiert', 'Support Vector Machines', 'Regression', 'Halbüberwachtes Lernen']
AC15	['Unstrukturiert', 'Autoencoding NN', 'Vereinfachen durch Dimension Reduction', 'Unüberwachtes Lernen']
AC16	['Strukturiert', 'Bayesian Methods', 'Regression', 'Halbüberwachtes Lernen']
AC17	['Strukturiert', 'Tree-Based Methods', 'Klassifikation', 'Überwachtes Lernen']
AC18	['Strukturiert', 'K-Nearest Neighbors ', 'Klassifikation', 'Überwachtes Lernen']
AC19	['Unstrukturiert', 'Neural Networks and Deep Learning', 'Klassifikation', 'Überwachtes Lernen']
AC20	['Strukturiert', 'Bayesian Methods', 'Klassifikation', 'Überwachtes Lernen']
AC21	['Strukturiert', 'Linear Models', 'Klassifikation', 'Überwachtes Lernen']
AC22	['Strukturiert', 'K-Nearest Neighbors ', 'Regression', 'Überwachtes Lernen']
AC23	['Strukturiert', 'Neural Networks and Deep Learning', 'Klassifikation', 'Überwachtes Lernen']
AC24	['Strukturiert', 'K-Nearest Neighbors ', 'Klassifikation', 'Halbüberwachtes Lernen']
AC25	['Strukturiert', 'Support Vector Machines', 'Regression', 'Überwachtes Lernen']
AC26	['Strukturiert', 'Bayesian Methods', 'Klassifikation', 'Halbüberwachtes Lernen']
AC27	['Strukturiert', 'Linear Models', 'Klassifikation', 'Halbüberwachtes Lernen']
AC28	['Strukturiert', 'Support Vector Machines', 'Klassifikation', 'Halbüberwachtes Lernen']
AC29	['Strukturiert', 'Support Vector Machines', 'Klassifikation', 'Überwachtes Lernen']
AC30	['Strukturiert', 'Neural Networks and Deep Learning', 'Regression', 'Überwachtes Lernen']
AC31	['Strukturiert', 'Linear Models', 'Regression', 'Überwachtes Lernen']
AC32	['Unstrukturiert', 'Neural Networks and Deep Learning', 'Regression', 'Überwachtes Lernen']
AC33	['Strukturiert', 'Bayesian Methods', 'Regression', 'Überwachtes Lernen']
AC34	['Strukturiert', 'SARSA', 'Interaktion statt bereitgestellten Daten', 'Bestärkendes Lernen']
AC35	['Strukturiert', 'Q - Learning', 'Interaktion statt bereitgestellten Daten', 'Bestärkendes Lernen']

Abbildung 7: Abgebildete Anwendungsmöglichkeiten von KI-Methoden (Eigene Darstellung)

	UM1	UM2	UM3	UM4	UM5	UM6	UM7	UM8	UM9	UM10
AC17	Matrix	Sehr Hoch	Hoch	Sehr Hoch	Gelabelt	Kategorisch	Hoch	Sehr Groß	nein	Gering
		Hoch	Normal	Hoch			Normal	Groß	nein	Mittel
		Normal	Gering	Normal			Gering	Neutral	nein	Hoch
								Klein	ja	

Abbildung 8: Beispiel von Kombinationen der Unterscheidungsmerkmale einer bestimmten Anwendungsmöglichkeit (Eigene Darstellung)

4.2 Entwicklung des Vorgehensmodells

4.2.1 Metalearning-Ansatz

In dieser Arbeit wurde der Ansatz in Anlehnung zu dem Themengebiet Metalearning eingerichtet (2.2, S. 4). In diesem Sinne wind ein Modell entwickelt in der Form eines Metalearning Systems, um eine automatische und systematische Benutzerführung bieten zu können, indem aus einer bestimmten Aufgabe eine geeignete KI-Methode vorschlägt (Brazdil *u. a.*, 2009, S. 2).

4.2.2 Erstellung von Trainingsdaten

In diesem Abschnitt werden die Phasen des CRISP-MD *Data Understanding*, *Data Preparation* und *Data Modeling* durchgeführt. In diesem Sinne wurden die 2470 manuell erstellten Daten als Analyse-Datensatz benannt und als *Richtig* gelabelt. Aus dieser Basis war das Ziel, durch Data Augmentation der Analyse-Datensatz zu vergrößern. Dafür wurde eine Oversampling-Technik als Strategie verwendet, nämlich der SMOTE Algorithmus (*Synthetic Minority Over-sampling Technique*)(Chawla *u. a.*, 2002). Der SMOTE-Algorithmus generiert synthetische Daten auf Basis von den Unterscheidungsmerkmalen aus klassifizierten gelabelten Daten. In einer vereinfachten Erklärung kann es gesagt werden, dass der Algorithmus Zusammenhänge und Abhängigkeiten in Referenzdaten erkennt und darauf aufbauend ähnliche Daten generiert. Bei diesen generierten Daten werden die Zusammenhänge und Abhängigkeiten zwischen den Unterscheidungsmerkmalen nach den Referenzdaten abgebildet.

Jedoch braucht der Algorithmus ein Gegenbeispiel, bei dem die Zusammenhänge und Abhängigkeiten zwischen den Unterscheidungsmerkmalen ganz anderes sind als die von den Referenzdaten. In anderen Worten braucht der Algorithmus auch falsche Daten. In diesem Sinne wurde aus den Unterscheidungsmerkmalen für die Charakterisierung der KI-Methoden (Abbildung 5, S. 10) ein Random-Datensatz mit Kombinationen aus den Beantwortungsmöglichkeiten des Kriterienrasters erstellt. Als Output dieser falschen Beantwortungen wurden die Anwendungsmöglichkeiten (Abbildung 7, S. 12) eingefügt. Sowohl die Kombinationen auf Basis des Kriterienrasters als auch die Anwendungsmöglichkeiten wurden zufällig in den Random-Datensatz eingefügt. Die Anzahl an Kombinationen in diesem Fall ist 24,2 Mio. Möglichkeiten. Dadurch dass die Anzahl an mögliche Kombinationen deutlich höher ist als die Anzahl an richtigen Kombinationen, wird angenommen, dass alle Anwendungsfälle aus dem Random-Datensatz falsch sind. In diesem Sinne wurden 250.000 Anwendungsfälle durch den Randomansatz generiert und als *Falsch* gelabelt.

Der Analyse-Datensatz und der Random-Datensatz wurden in einen Datensatz zusammengefügt, der als SMOTE-Datensatz benannt wird. Die Datenaufteilung in den Klassen Analyse und Random ist in der Abbildung 9 auf der linken Seite dargestellt. Der SMOTE Algorithmus wurde dann verwendet, um mit synthetischen Daten die Analyse-Klasse (Klasse der richtigen Daten) zu erweitern. Nach der Ausführung des SMOTE-Algorithmus hat die Klasse Analyse die gleiche Anzahl an Daten als die Klasse Random (250.000 Daten). Die Neue Aufteilung des SMOTE-Datensatzes ist in der Abbildung 9 auf der rechten Seite dargestellt.

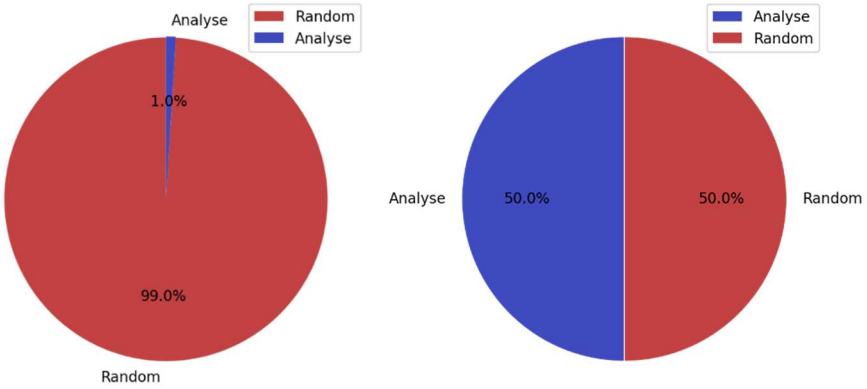

Abbildung 9: Datenaufteilung des SMOTE-Datensatzes vor (links) und nach (rechts) Anwendung des SMOTE-Algorithmus (Eigene Darstellung)

Aus dem SMOTE-Datensatz wurden nur die Analysedaten als Trainingsdaten für das Modell verwendet und wird als Training-Datensatz bezeichnet (Die Daten aus dem Analyse-Datensatz und die aus dem SMOTE-Algorithmus generierten Daten). Der Training-Datensatz enthält 250.000 Trainings-Beispiele, bei denen als Zielgröße die Anwendungsmöglichkeiten (Abbildung 7, S. 12) definiert worden sind. Aus dem Training-Datensatz wurde 85% (212.500 Beispiele) für das Lernverfahren und 15% (37.500 Beispiele) für das Testen verwendet. Im folgenden Abschnitt wird erklärt wie die Erstellung des Metalearning-Modells

4.2.3 Vorgehensmodell – Training des Metalearning-Modells

Das Vorgehensmodell besteht aus der Entwicklung eines Metalearning-Modells. Dafür wurden zwei Algorithmen verwendet, die zu der KI-Methodenfamilie *Tree-Based Methods* oder baumbasierten Methoden gehören. Beide verwenden das Gradient Boosting Framework und weisen gute Performance und Skalierbarkeit auf (*LightGBM Documentation*, 2020; *XGBoost Documentation*, 2020). Als Vorteil der Auswahl dieser KI-Methodenfamilie steht die gute Nachvollziehbarkeit des Modells (Schacht und Lanquillon, 2019), aus dem ein Entscheidungsbaum mit den verwendeten Features entworfen werden kann. Darüber Hinaus ist durch die Anwendung dieser Methoden möglich, die Wichtigkeit der Features für die Entscheidung auszuwerten, was für das Wissensaufbau bezüglich der Domäne des Modells hilfreich ist.

Die Entscheidung für die Erstellung von Zwei Metalearning-Modelle statt nur eins ist Zwecks Plausibilisierung. Die beide Verfahren sollten ähnliche Vorschläge liefern, da das Lernverfahren jedes Modell auf den gleichen Daten basiert. Aus dieser Hypothese können die Ergebnisse besser beurteilt werden und somit die Fehleranfälligkeit minimieren (Pedro Domingos, 2012, S. 85).

In diesem Sinne haben die beide Modelle aus den gleichen Daten gelernt. Für die Performanceevaluation wurden die Genauigkeit oder *Precision*, Recall und F1-Score ausgewählt. Das Ergebnis ist in der Abbildung 10 dargestellt. Kein großer Unterschied zwischen den Performances der beiden Modelle wurde beobachtet. Die durchschnittliche Performance bei dem Test-Daten von ca. 0,96 wurde bei den drei Performancekennzahlen und bei den zwei Modellen beobachtet, was eine gute Performancestabilität aufweist.

precision report:
XGBoost Model: 0.958
LightGBM Model: 0.958
recall report:
XGBoost Model: 0.958
LightGBM Model: 0.957
f1_score report:
XGBoost Model: 0.958
LightGBM Model: 0.957

Abbildung 10: Performanceevaluation des Vorgehensmodells (Eigene Darstellung)

Darüber hinaus wurden die Modelle mit den jeweiligen wichtigsten Merkmalen auf Basis der Kennzahl *Total Gain* in der Abbildung 11 dargestellt. Aus den 40 Merkmalen ist die Auflistung bis auf den vierten Platz gleich bei den beiden Modellen und unter den zehn wichtigsten Merkmalen aus den beiden Modellen sind nur zwei Unterschiede bei jeder Auflistung zu beobachten. Diese Erkenntnis zeigt, dass die Ergebnisse der Modelle plausibel und kongruent zu einander sind. Ein weiterer Schritt hin zur Evaluation wird Anhand ein Praxisbeispiel erläutert.

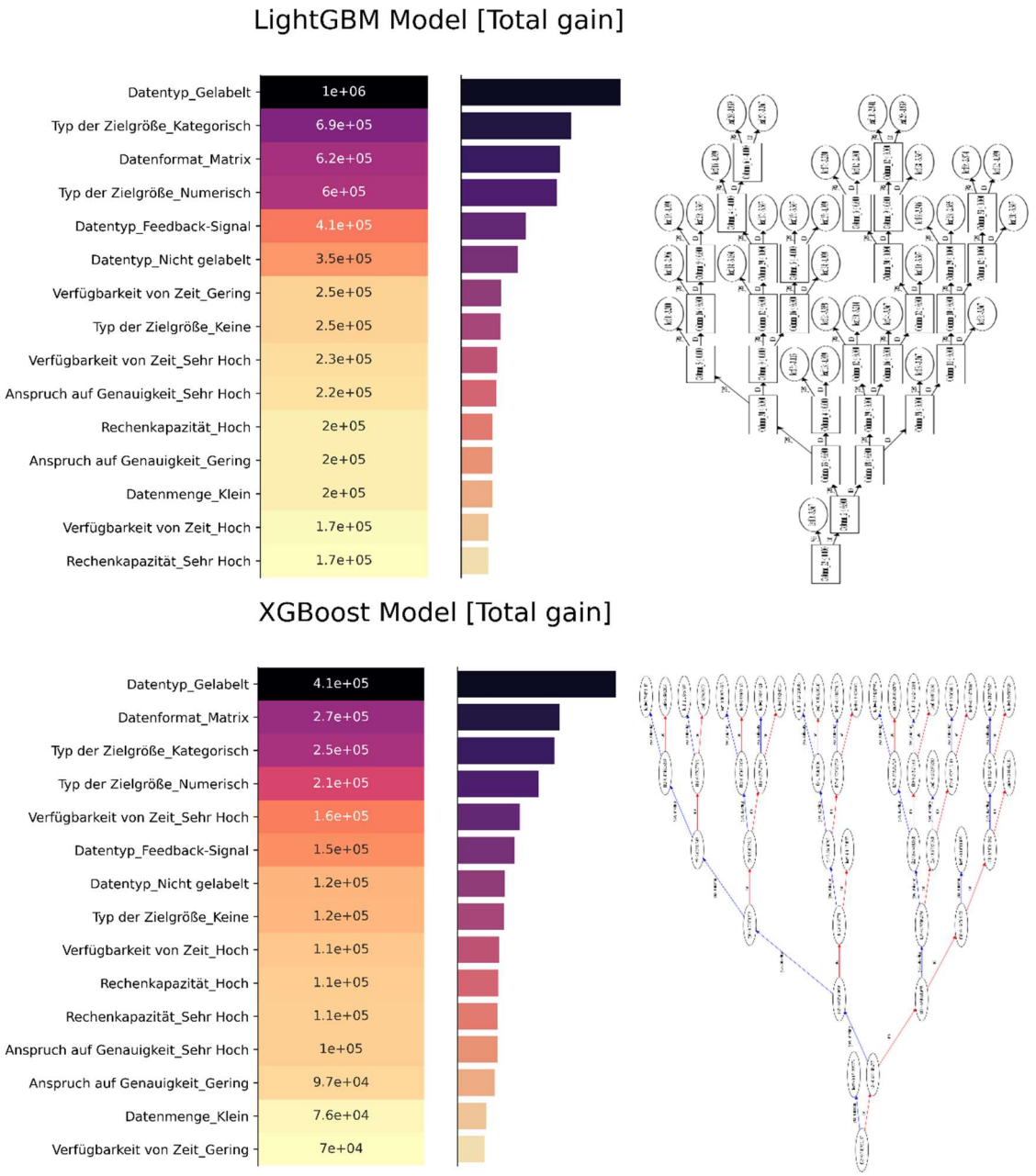

Abbildung 11: Wichtigste Merkmale und Modelldarstellung aus den beiden Metalearning-Modellen (Eigene Darstellung)

4.3 Fallbeispiel im Stammdatenmanagement im Rahmen von CRISP-DM

Für die weitere Evaluation des Vorgehensmodells wurden aus der Funktionsarchitektur des Stammdatenmanagements (Otto, Hüner und Österle, 2012, S. 418–422) drei Aufgaben ausgewählt. Darauf aufbauend wurden drei Probleme abgebildet, die durch die Anwendung des Vorgehensmodells gelöst wurden. Das Ziel dieser Evaluation ist eine Simulation von einer Praxisanwendung durchzuführen, um zu testen, wie das Modell in der Praxis performt.

4.3.1 Problem 1

1. **Funktionsgruppe**: Metadaten Management und Stammdatenmodellierung

2. **Aufgabegruppe**: Model Analysis
3. **Zielsetzung:** Beziehungen zwischen Datentypen sollen automatisch erkannt werden
4. **Storytelling:**

Für dieses Fallbeispiel, wird angenommen, dass die Daten in der Form einer Matrix ist und die Datenqualität ist hochgeschätzt. Der Mitarbeiter ist nicht besonderes unter Druck, deswegen hat er eine normale Zeitverfügbarkeit. Dadurch, dass die Daten hochqualitativ geschätzt sind, ist der Anspruch auf Genauigkeit Hoch. Die Daten sind nicht gelabelt und der Mitarbeiter möchte die Möglichkeiten der Daten explorieren. Er verfügt ein normales Arbeitsrechner, deswegen schätzt er, dass die Rechenkapazität normal ist. Die Datenmenge ist groß, es geht um keine Folge von Entscheidungen und die Anzahl an Dimensionen liegt bei ca. 30, was er als „Mittel" schätzt.

5. **Aus dem Storytelling ausgefülltes Kriterienraster.**

Datenformat	Datenqualität	Verfügbarkeit von Zeit	Anspruch auf Genauigkeit	Datentyp	Typ der Zielgröße	Rechenkapazität	Datenmenge	Folge von Entscheidungen	Anzahl an Dimensionen (Features)
Matrix	Hoch	Normal	Hoch	Nicht gelabelt	Keine	Normal	Groß	nein	Mittel

Abbildung 12: Kriterienraster aus Problem 1 (Eigene Darstellung)

6. **Vorschläge aus den Modellen**

Vorschlag	Light GBM
1	['Strukturiert', 'K-Means', 'Muster erkennen und Clustering', 'Unüberwachtes Lernen']
2	['Strukturiert', 'PCA', 'Vereinfachen durch Dimension Reduction', 'Unüberwachtes Lernen']
Vorschlag	**XGBoost**
1	['Strukturiert', 'PCA', 'Vereinfachen durch Dimension Reduction', 'Unüberwachtes Lernen']
2	['Strukturiert', 'K-Means', 'Muster erkennen und Clustering', 'Unüberwachtes Lernen']

Abbildung 13: Vorschläge aus den Modellen – Problem 1 (Eigene Darstellung)

7. **Kommentare zu den Vorschlägen**

Das Modell hat erkannt, dass es um eine strukturierte Datensatz geht, und dass die Lernform unüberwachtes Lernen angewendet werden sollte. Dadurch, dass kein Plan existiert hat, sind zwei Vorgehensweisen vorgeschlagen, erstens die Vereinfachung durch die *Dimension Reduction* oder Dimensionsreduzierung, um die Daten einfacher darzustellen und explorieren dazu war die Empfehlung die PCA als KI-Methode anzuwenden, da die Rechenkapazität nur normal war und die Zeitverfügbarkeit auch normal war. Zweitens die Clustering der Daten, um Zusammenhänge zu erkennen und darauf aufbauend auch die Daten weiter zu explorieren, dazu wurde die KI-Methode K-Means vorgeschlagen, da die Datenmenge groß war und die Zeitverfügbarkeit für komplexere Modelle gefehlt hat.

4.3.2 Problem 2

1. **Funktionsgruppe**: Qualitätsmanagement für Stammdaten.
2. **Aufgabegruppe**: Datenbereinigung
3. **Zielsetzung**: Es soll sichergestellt werden, dass keine ungültigen Daten in Unternehmenssysteme eingegeben werden.
4. **Storytelling**:

Bei diesem Beispiel hat der Mitarbeiter erkannt, dass manchmal ungültige Daten in das System eingetragen werden. Das Projekt soll nachhaltig sein, das bedeutet, dass der Anspruch auf Genauigkeit vor der Verfügbarkeit der Zeit im Kriterienraster kommt. Die gelabelten Daten, auf deren Basis das Modell lernen wird enthält hohen Grad an Qualität und die Zielgröße ist numerisch. Der Mitarbeiter verfügt Ressourcen in einem Cluster im Data Center des Unternehmens die Anzahl an Features ist normal und keine Folge von Entscheidungen getroffen werden soll.

5. **Aus dem Storytelling ausgefülltes Kriterienraster.**

Datenformat	Datenqualität	Verfügbarkeit von Zeit	Anspruch auf Genauigkeit	Datentyp	Typ der Zielgröße	Rechenkapazität	Datenmenge	Folge von Entscheidungen	Anzahl an Dimensionen (Features)
Matrix	Hoch	Normal	Sehr Hoch	Gelabelt	Numerisch	Hoch	Groß	nein	Mittel

Abbildung 14: Kriterienraster aus dem Problem 2 (Eigene Darstellung)

6. **Vorschläge aus den Modellen.**

Vorschlag	Light GBM
1	['Strukturiert', 'Tree-Based Methods', 'Regression', 'Überwachtes Lernen']
2	['Strukturiert', 'Support Vector Machines', 'Regression', 'Überwachtes Lernen']
Vorschlag	**XGBoost**
1	['Strukturiert', 'Tree-Based Methods', 'Regression', 'Überwachtes Lernen']
2	['Strukturiert', 'Support Vector Machines', 'Regression', 'Überwachtes Lernen']

Abbildung 15: Vorschläge aus den Modellen – Problem 2 (Eigene Darstellung)

7. **Kommentare aus den Vorschlägen**

Es wurde erkannt, dass die Daten strukturiert sind, weil die Trainingsdaten gelabelt waren. Dadurch dass die Zielgröße vorhanden sind ist die Lernform überwachtes Lernen. Die Regression wird in dem Fall vorgeschlagen, weil die Zielgröße numerisch ist. Als erste Empfehlung wurden in beiden Fällen die *Tree-Based-Algerithms* erst vorgeschlagen, da die Anspruch auf die Genauigkeit sehr hoch ist und die Verfügbarkeit von Zeit nicht so hoch war, was bei den Support Vector Machines eine Herausforderung darstellen könnte.

4.3.3 Problem 3

1. **Funktionsgruppe**: Qualitätsmanagement für Stammdaten.
2. **Aufgabegruppe**: Datenbereinigung
3. **Zielsetzung**: Identifiziert bestimmte Muster in Daten-Repositories.
4. **Storytelling**:

Der Mitarbeiter arbeitet indirekt mit Stammdatenmanagement und hat andere wichtige Projekte am Laufen. Er hat einen kleinen Datensatz zur Verfügung, bei dem er schnell ein Muster entdecken möchte, aber bevor er zu viel Zeit investiert, will er erstens mit einem explorativen Ansatz Vorgehen. Die Daten sind in der Form einer Matrix im Excel. Die Datenqualität ist normal, die Verfügbarkeit von Zeit ist gering und der Anspruch auf Genauigkeit auch. Der Datentyp ist gelabelt und er verfügt einen normalen Computer. Die Datenmenge ist neutral, nicht groß, aber auch nicht zu klein. Es besteht keine Folge von Entscheidungen und die Anzahl an Dimensionen ist bei 25 von ihm als Mittel bezeichnet.

5. Aus dem Storytelling ausgefülltes Kriterienraster.

Datenformat	Datenqualität	Verfügbarkeit von Zeit	Anspruch auf Genauigkeit	Datentyp	Typ der Zielgröße	Rechenkapazität	Datenmenge	Folge von Entscheidungen	Anzahl an Dimensionen (Features)
Matrix	Normal	Gering	Gering	Gelabelt	Kategorisch	Normal	Neutral	nein	Mittel

Abbildung 16: Kriterienraster aus dem Problem 3 (Eigene Darstellung)

6. Vorschläge aus den Modellen

Vorschlag	Light GBM
1	['Strukturiert', 'K-Nearest Neighbors ', 'Klassifikation', 'Überwachtes Lernen']
2	['Strukturiert', 'Linear Models', 'Klassifikation', 'Überwachtes Lernen']
Vorschlag	**XGBoost**
1	['Strukturiert', 'K-Nearest Neighbors ', 'Klassifikation', 'Überwachtes Lernen']
2	['Strukturiert', 'Linear Models', 'Klassifikation', 'Überwachtes Lernen']

Abbildung 17: Vorschläge aus den Modellen – Problem 3 (Eigene Darstellung)

7. Kommentare aus den Vorschlägen.

Das Vorgehensmodell hat erkannt, dass die Daten strukturiert in der Form einer Matrix sind. Dadurch, dass eine Zielgröße vorhanden war und ihr Typ kategorisch war, ging es um Überwachtes Lernen und die Klassifikation als Methode zum Ziel. Dadurch, dass die Daten Neutral waren und wegen der geringen verfügbaren Zeit und niedrigeren Anspruch auf Genauigkeit wurden einfachere KI-Methoden vorgeschlagen, nämlich K-Nearest Neighbors und Linear Models.

5 Diskussion und Ausblick

In dieser Arbeit wurde basierend auf der Design Science Research Methodik ein Artefakt in der Form eines Vorgehensmodells zur Beantwortung der Forschungsfrage entwickelt. Das Vorgehensmodell hat sich auf die Idee von Metalearning bezogen, indem die Zielsetzung war, die praktische Umsetzung von KI-Methoden zu unterstützen. In diesem Sinne ist das entworfene Artefakt fähig, aus Charakteristiken der Aufgabestellung, passende KI-Methoden vorzuschlagen.

Dafür wurde erstens ein Kriterienraster zur Charakterisierung des zu lösenden Problems entworfen. Zweitens wurde ein Datensatz mit Fallbeispielen von richtigen KI-Methodenauswahlen erstellt. Darauf aufbauend wurden zwei baumbasierten Modellen erstellt, die aus den erstellen Daten gelernt haben. Diese Modelle wurden auf Basis von Simulationen von Aufgabenstellungen aus Stammdatenmanagement getestet. Es wurde beobachtet, dass die vorgeschlagenen Ergebnisse übereinstimmend mit der Theorie waren, somit wurde behauptet, dass das Modell eine gute Basis für weitere Forschungen darstellt.

Weitere Plausibilisierungen sollten durchgeführt werden, um die Eignung des Modells in Praxissituationen nachzuprüfen. Außerdem, könnte als Verbesserung weitere Beispiele von richtigen Anwendungen von KI-Methoden für das Training des Modells genutzt werden, somit weitere Fälle abgedeckt werden. Eine Vertiefung aus der KI-Methodenfamilien in die KI-Algorithmen hätte auch Forschungspotenzial, sowie die Erweiterung der Eingabemerkmale für die Abdeckung von weiteren Strategien, wie Transfer Learning und Data Augmentation, oder unterschiedlichen Vorgehensweisen wie Batch Learning und Online Learning.

6 Literatur

Ali, R., Lee, S. und Chung, T. C. (2017) „Accurate multi-criteria decision making methodology for recommending machine learning algorithm“, *Expert Systems with Applications*. Elsevier Ltd, 71, S. 257–278. doi: 10.1016/j.eswa.2016.11.034.

Braschler, M., Staderlmaan, T. und Stockinger, K. (2019) *Analytics & Applied Data Science*. Verfügbar unter: https://brightcape.nl/analytics-applied-data-science.

Brazdil, P. *u. a.* (2009) *Metalearning Applications to Data Mining*, *Metalearning*. Springer-Verlag Berlin Heidelberg. doi: 10.1007/978-3-540-73263-1.

Caruana, R. und Niculescu-Mizil, A. (2006) „An empirical comparison of supervised learning algorithms“, *ACM International Conference Proceeding Series*, 148, S. 161–168. doi: 10.1145/1143844.1143865.

Chawla, N. V. *u. a.* (2002) „SMOTE: Synthetic Minority Over-sampling Technique Nitesh“, *Journal of Artificial Intelligence Research*. doi: 10.1613/jair.953.

Das, T. kumar und Mishra, M. R. (2011) „A Study on Challenges and Opportunities in Master Data Management“, *International Journal of Database Management Systems*, 3(2), S. 129–139. doi: 10.5121/ijdms.2011.3209.

Fettke, P. (2006) „State-of-the-art des state-of-the-art: Eine Untersuchung der Forschungsmethode ‚Review‘ innerhalb der Wirtschaftsinformatik“, *Wirtschaftsinformatik*, 48(4), S. 257–266. doi: 10.1007/s11576-006-0057-3.

Hastie, T., Tibshirani, R. und Friedman, J. (2009) *The elements of statistical learning*.

Joshi, A. V (2020) *Machine Learning and Artificial Intelligence an Introduction*. doi: https://doi.org/10.1007/978-3-030-26622-6.

Kreutzer, R. T. und Sirrenberg, M. (2020) *Understanding Artificial Intelligence*, *Encore*. doi: 10.1007/978-3-030-25271-7.

LightGBM Documentation (2020) *LightGBM*. Verfügbar unter: https://lightgbm.readthedocs.io/en/latest/ (Zugegriffen: 19. Juli 2020).

Madlberger, M. (2011) „Can data quality help overcome the penguin effect? The case of item master data pools“, *19th European Conference on Information Systems, ECIS 2011*.

Nandy, A. und Biswas, M. (2018) *Reinforcement Learning: With Open AI, TensorFlow and Keras Using Python. Apress*. doi: 10.1007/978-1-4471-4285-0_11.

Otto, B. und Hüner, K. M. (2009) „Funktionsarchitektur für unternehmensweites Stammdatenmanagement“, *Universität St. Gallen*, (January 2015), S. 68. Verfügbar unter: http://intranet.iwi.unisg.ch/org/iwi/iwi_pub.nsf/wwwPublAuthorGer/E63C7A6DAF0A3B43C12576480022E057/$file/MdmFunktionsarchitektur.pdf.

Otto, B., Hüner, K. M. und Österle, H. (2012) „Toward a functional reference model for master data quality management“, *Information Systems and e-Business Management*, 10(3), S. 395–425. doi: 10.1007/s10257-011-0178-0.

Pedro Domingos (2012) „A Few Useful Things to Know About Machine Learning“, *communications of the ACM*, 55(10), S. 79–88.

Peffers, K. *u. a.* (2007) „A design science research methodology for information systems research“, *Journal of Management Information Systems*, 24(3), S. 45–77. doi: 10.2753/MIS0742-1222240302.

Samuel, A. L. (1959) „Some studies in machine learning using the game of checkers“, *IBM Journal of Research and Development*. doi: 10.1147/rd.441.0206.

Schacht, S. und Lanquillon, C. (2019) *Blockchain und maschinelles Lernen*, *Blockchain und maschinelles Lernen*. doi: 10.1007/978-3-662-60408-3.

Shearer, C. *u. a.* (2000) „The CRISP-DM model: The New Blueprint for Data Mining“, *Journal of Data Warehousing*, 5(4), S. 13–22. Verfügbar unter: www.spss.com%5Cnwww.dw-institute.com.

Webster, J. und Watson, R. T. (2002) „Analyzing the Past to Prepare for the Future: Writing a Literature Review.“, *MIS Quarterly*, 26(2), S. xiii–xxiii. doi: 10.1.1.104.6570.

Wolpert, D. H. (1996) „The Lack of a Priori Distinctions between Learning Algorithms“, *Neural Computation*, 8(7), S. 1341–1390. doi: 10.1162/neco.1996.8.7.1341.

XGBoost Documentation (2020) *XGBoost*. Verfügbar unter: https://xgboost.readthedocs.io/en/latest/ (Zugegriffen: 19. Juli 2020).

Zschech, P. *u. a.* (2018) „Data Science Skills and Enabling Enterprise SystemsData Science Skills and Enabling Enterprise Systems“, *HMD Praxis der Wirtschaftsinformatik*, 55(1), S. 163–181. doi: 10.1365/s40702-017-0376-4.

Zschech, P. *u. a.* (2020) „Intelligent User Assistance for Automated Data Mining Method Selection“, *Business and Information Systems Engineering*, 62(3), S. 227–247. doi: 10.1007/s12599-020-00642-3.