



**Masterarbeit**  
**im Studiengang Computer Science and Media**

**KONZEPTION UND IMPLEMENTIERUNG EINES  
EMPFEHLUNGSSYSTEMS FÜR DIE AUSWAHL UND ANWENDUNG  
VON XAI-METHODEN**

vorgelegt von

**Verena Barth**

an der Hochschule der Medien

am 02.07.2021

zur Erlangung des akademischen Grades eines  
Master of Science

**Erstprüfer:** Prof. Dr. Christian Becker-Asano  
**Zweitprüfer:** Dr. Frank Köhne

# Eidesstattliche Erklärung

Hiermit versichere ich, Verena Barth, ehrenwörtlich, dass ich die vorliegende Masterarbeit mit dem Titel: „Konzeption und Implementierung eines Empfehlungssystems für die Auswahl und Anwendung von XAI-Methoden“ selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden. Ich habe die Bedeutung der ehrenwörtlichen Versicherung und die prüfungsrechtlichen Folgen einer unrichtigen oder unvollständigen ehrenwörtlichen Versicherung zur Kenntnis genommen.

Datum: \_\_\_\_\_ Unterschrift: \_\_\_\_\_

# Gender Erklärung

Aus Gründen der besseren Lesbarkeit wird in dieser Masterarbeit die Sprachform des generischen Maskulinums verwendet. Es wird an dieser Stelle darauf hingewiesen, dass diese ausschließliche Verwendung der männlichen Form geschlechtsunabhängig verstanden werden soll.

# Kurzfassung

Um den Einsatz von Explainable AI (XAI) und dadurch die Nachvollziehbarkeit komplexer KI-Systeme zu fördern, wird der Frage nachgegangen, wie man den Nutzer bei der Auswahl und der anschließenden Anwendung von geeigneten XAI-Methoden auf Black-Box Modelle durch Empfehlungen unterstützen kann. Gemäß dem Requirements Driven Design Science Research Framework wird dafür, nach einer initialen Anforderungsanalyse in Form von qualitativen, halbstrukturierten Interviews, ein webbasiertes XAI-Empfehlungssystem (XAIR) konzipiert und implementiert. Der XAIR schlägt, basierend auf den Eingaben des Nutzers unter Berücksichtigung seiner Präferenzen, die am besten geeigneten, modellagnostischen und post-hoc anwendbaren XAI-Methoden vor. Dafür wird zunächst geklärt, wie die Eignung von XAI-Methoden beurteilt werden kann, welche Kriterien eines spezifischen Modell-, Daten- und Nutzungskontexts diese beeinflussen und wie sich dieses Wissen formalisieren lässt. Aufgrund der Vagheit des vorhandenen XAI-Expertenwissens ist in diesem Kontext die Fuzzy-Logik von besonderer Relevanz. Die resultierende, modular aufgebaute Software kann als Web-Anwendung oder automatisiert, bspw. innerhalb einer ML-Pipeline, verwendet werden. Eine Nutzerevaluation weist darauf hin, dass der entwickelte XAIR durch eine begründete Empfehlung nicht nur die Auswahl einer XAI-Methode und ihrer Implementierung vereinfacht, sondern darüber hinaus durch Bereitstellung tieferer und gezielter Informationen diesbezüglich auch die Bereitschaft der tatsächlichen Anwendung erhöht.

## Abstract

In order to encourage the use of Explainable AI (XAI) and thereby improve the interpretability of complex AI systems, the question of how users can be assisted when selecting and subsequently applying suitable XAI methods is investigated. To help a user select an appropriate XAI method for black-box models by means of recommendations, a web-based XAI recommendation system (XAIR) is designed and implemented. Based on the user's inputs incorporating their preferences, the most appropriate model-agnostic and post-hoc applicable XAI methods are suggested. According to the Requirements Driven DSR framework an initial requirements analysis in the form of qualitative, semi-structured interviews is conducted. To identify and formalize context-specific criteria influencing the applicability of an XAI method, a measurement for its suitability is determined. Due to the vagueness of the existing XAI expert knowledge, fuzzy logic is of particular relevance in this context. The modular software resulting from this work can be further extended with XAI methods and other criteria. Furthermore, it can be used standalone as a web application or automated within an ML pipeline. A user evaluation indicates that the XAIR simplifies the selection of an XAI method and its implementation by providing a well-founded recommendation. Moreover, it also increases the motivation for the user to actually apply the XAI method by outlining more targeted, in-depth information about the respective method.

# Inhaltsverzeichnis

<b>Abbildungsverzeichnis</b>	<b>vi</b>
<b>Tabellenverzeichnis</b>	<b>vii</b>
<b>Abkürzungsverzeichnis</b>	<b>viii</b>
<b>1. Einleitung</b>	<b>1</b>
1.1. Problemstellung . . . . .	1
1.2. Zielsetzung . . . . .	2
1.3. Methodik und Aufbau der Arbeit . . . . .	2
<b>2. Expertensystem für die Auswahl einer XAI-Methode</b>	<b>5</b>
2.1. Machine Learning und seine Black-Box Problematik . . . . .	5
2.2. Explainable Artificial Intelligence (XAI) . . . . .	6
2.2.1. Taxonomie interpretierbarer Methoden . . . . .	7
2.2.2. Ist-Situation und verwandte Arbeiten . . . . .	8
2.3. Empfehlungssysteme und Expertensysteme . . . . .	11
2.4. Fuzzy-Expertensysteme . . . . .	12
2.4.1. Fuzzy-Logik . . . . .	12
2.4.2. Anwendung der Fuzzy-Logik in einem Expertensystem . . . . .	15
<b>3. Ziele und Anforderungen eines XAI-Empfehlungssystems</b>	<b>17</b>
3.1. Ableitung der Forschungsfragestellung . . . . .	17
3.2. Ziele und Annahmen des XAI-Empfehlungssystems . . . . .	17
3.3. Anforderungsanalyse . . . . .	18
3.3.1. Aufbau und Durchführung . . . . .	18
3.3.2. Erhobener Anforderungskatalog . . . . .	19
3.3.3. Anforderungsübersicht . . . . .	22
<b>4. Konzeption des XAI-Empfehlungssystems</b>	<b>24</b>
4.1. Identifikation geeigneter XAI-Methoden . . . . .	24
4.1.1. Beurteilung der Eignung einer XAI-Methode . . . . .	24
4.1.2. Eingrenzung der XAI-Methoden und Auswahl der Implementierungen . . . . .	25
4.1.3. Voraussetzungen für die Anwendung der XAI-Methoden . . . . .	32
4.1.4. Ableitung von Kriterien zur Beurteilung der Eignung von XAI-Methoden . . . . .	33
4.2. Auswahl einer Wissensrepräsentation . . . . .	38
4.2.1. Problem der Erhebung der Kriterienwerte und ihrer Auswirkungen . . . . .	38
4.2.2. Wahl der zu verwendenden Logik und Wissensrepräsentation . . . . .	39
4.3. Konfiguration des Fuzzy-Expertensystems . . . . .	39
4.3.1. Definition der Ein- und Ausgaben . . . . .	39

4.3.2.	Erhebung und Ermittlung der Systemeingaben . . . . .	41
4.3.3.	Identifikation der Eingaben des Fuzzy-Systems . . . . .	49
4.3.4.	Festlegung der Fuzzifizierung und Defuzzifizierung . . . . .	49
4.3.5.	Erstellung der Regelbasis . . . . .	54
4.3.6.	Aufbau der grafischen Benutzeroberfläche (GUI) . . . . .	57
<b>5.</b>	<b>Implementierung des XAIRs</b>	<b>60</b>
5.1.	Verwendete Technologien . . . . .	60
5.2.	Ausgewählte Implementierungsdetails . . . . .	60
5.2.1.	Konfiguration . . . . .	60
5.2.2.	Erweiterbarkeit hinsichtlich neuer Kriterien und XAI-Methoden . . . . .	61
5.2.3.	Validierung der Eingabeparameter . . . . .	63
5.2.4.	Nachvollziehbarkeit der Empfehlungsergebnisse . . . . .	63
<b>6.</b>	<b>Evaluation</b>	<b>64</b>
6.1.	Aufbau und Durchführung . . . . .	64
6.2.	Ergebnisse der Nutzerevaluation . . . . .	65
6.3.	Diskussion der Resultate . . . . .	68
6.3.1.	Erfüllung der Erwartungshaltung der Nutzer . . . . .	68
6.3.2.	Erkenntnisse bezüglich der Forschungsfrage . . . . .	69
6.3.3.	Einordnung des Ergebnisses in den Stand der Forschung . . . . .	70
6.3.4.	Limitationen und Herausforderungen . . . . .	71
<b>7.</b>	<b>Fazit und Ausblick</b>	<b>74</b>
7.1.	Fazit . . . . .	74
7.2.	Ausblick . . . . .	76
	<b>Literaturverzeichnis</b>	<b>I</b>
<b>A.</b>	<b>Anhang</b>	<b>A1</b>
A.1.	Leitfaden des Fragenkatalogs zur Anforderungsanalyse . . . . .	A1
A.2.	Implementierungsbewertung der ausgewählten XAI-Methoden . . . . .	A2
A.3.	Formulierungen bei der Eingabe benötigter Parameter . . . . .	A6
A.4.	Empfehlung zur Ausführung der Preprocessing Operationen . . . . .	A9
A.5.	Vergleich der Korrelationskoeffizienten $\Phi_K$ und Pearson's $\rho$ . . . . .	A10
A.6.	Beispiele der automatisierten Datenanalyse . . . . .	A12
A.7.	Übersicht der Fuzzy-Regeln . . . . .	A18
A.8.	Screenshots der GUI der Web-Anwendung . . . . .	A20
A.9.	Anwendungsfall der Evaluation . . . . .	A27
A.10.	Leitfaden des Fragenkatalogs zur Evaluation . . . . .	A28

# Abbildungsverzeichnis

2.1.	Klassische und Fuzzy-Mengen . . . . .	13
2.2.	Eigenschaften und Operationen von Fuzzy-Mengen . . . . .	14
2.3.	Schematischer Aufbau eines Fuzzy-Expertensystems . . . . .	15
2.4.	Max-Min-Inferenz und Defuzzifizierung . . . . .	16
4.1.	Schritte und Formatbezeichnungen des Preprocessings . . . . .	32
4.2.	Übersicht eignungsbeeinflussender Kriterien des Modell-, Daten- und Nutzungskontexts	37
4.3.	Use Case Diagramm des Empfehlungserhalts . . . . .	41
4.4.	Zustandsdiagramm der Konfiguration einer ML-Pipeline . . . . .	42
4.5.	Grafische Ausgabe des $g_k$ am Beispiel der Features des UCI Adult Datensatzes . . . . .	45
4.6.	EqualWidth Binning der Features „Age“ und „Capital Gain“ (UCI Adult Datensatz) .	48
4.7.	Visualisierung der durch <i>scikit-fuzzy</i> verfälschten Zugehörigkeitsfunktionen der Tabelle 4.4	51
4.8.	Zugehörigkeitsfunktionen der Ausgabevariablen des Fuzzy-Expertensystems . . . . .	53
4.9.	Beispielhafte Ergebnisse betrachteter Defuzzifizierungsstrategien . . . . .	53
5.1.	Schematische Komponentenarchitektur des XAI-Empfehlungssystems . . . . .	62
A.1.	Durchschnittliche Differenz von $\Phi_K$ und Pearson's $\rho$ (50 Iterationen, 1000 Instanzen) .	A11
A.2.	Durchschnittliche Differenz von $\Phi_K$ und Pearson's $\rho$ (500 Iterationen, 10000 Instanzen)	A11
A.3.	Summe der Feature-Korrelationszugehörigkeiten pro Fuzzy-Menge (UCI Adult Datensatz)	A13
A.4.	Summe der Feature-Korrelationszugehörigkeiten pro Fuzzy-Menge (Titanic Datensatz) .	A14
A.5.	Summe beispielhafter Feature-Korrelationszugehörigkeiten pro Fuzzy-Menge . . . . .	A15
A.6.	EqualWidth Binning des Features „Hours per Week“ (UCI Adult Datensatz) . . . . .	A16
A.7.	Screenshot Ergebnisseite . . . . .	A20
A.8.	Screenshot Eingabeseite . . . . .	A21
A.9.	Screenshot Erklärungsseite . . . . .	A22
A.10.	Screenshot Detailseite der empfohlenen XAI-Methode (ALE) . . . . .	A23
A.11.	Screenshot Detailseite der empfohlenen Implementierung (ALE) . . . . .	A24
A.12.	Screenshot Allgemeine Empfehlungsseite . . . . .	A25
A.13.	Screenshot FAQ-Seite . . . . .	A26

# Tabellenverzeichnis

4.1.	Voraussetzungen für die Anwendung der XAI-Methoden . . . . .	33
4.2.	Daumenregeln der Interpretation von Korrelationskoeffizienten . . . . .	46
4.3.	Bildung des Gesamtkorrelationswertes durch die fuzzy OR-Verknüpfung . . . . .	46
4.4.	Definition der Zugehörigkeitsfunktionen mit Dezimalzahlen bei <i>scikit-fuzzy</i> . . . . .	50
4.5.	Zugehörigkeitsfunktionen der Eingabevariablen des Fuzzy-Expertensystems . . . . .	52
4.6.	Bewertung der XAI-Methoden hinsichtlich der Voraussetzungen . . . . .	54
4.7.	Bewertung der XAI-Methoden bzgl. der fuzzy Kriterien . . . . .	55
A.1.	Implementierungsvergleich PDP+ICE . . . . .	A3
A.2.	Implementierungsvergleich ALE . . . . .	A3
A.3.	Implementierungsvergleich SHAP . . . . .	A4
A.4.	Implementierungsvergleich Anchors . . . . .	A4
A.5.	Implementierungsvergleich Counterfactual Explanations . . . . .	A5
A.6.	Implementierungsvergleich PFI . . . . .	A5
A.7.	Formulierungen der Erfragung der Ausschlusskriterien . . . . .	A6
A.8.	Formulierungen der Erfragung eignungsbeeinflussender Eingabeparameter . . . . .	A7
A.9.	Deskriptive Statistik der Verteilungen einer Iteration . . . . .	A10
A.10.	Zuordnung der $g_K$ Werte zu Korrelationszugehörigkeitsmengen (UCI Adult Datensatz)	A13
A.11.	Zuordnung der $g_K$ Werte zu Korrelationszugehörigkeitsmengen (Titanic Datensatz) . .	A14
A.12.	Zuordnung beispielhafter Korrelationswerte zu Korrelationszugehörigkeitsmengen . . . .	A15
A.13.	Die Diskretisierbarkeit betreffende Charakteristika der Features . . . . .	A17



# Abkürzungsverzeichnis

AGG	Allgemeines Gleichbehandlungsgesetz.
ALE	Accumulated Local Effects.
BSI	Bundesamt für Sicherheit in der Informationstechnik.
CF	Counterfactual Explanations.
CFProto	Counterfactuals guided by Prototypes.
COG	Center of Gravity.
DNN	Tiefes Neuronales Netz (Deep Neural Network).
DSGVO	Datenschutzgrundverordnung.
FOI	Features of Interest.
FOM	First of Maximum.
GUI	Grafische Benutzeroberfläche (Graphical User Interface).
HCI	Mensch-Computer-Interaktion (Human-Computer Interaction).
ICE	Individual Conditional Expectation.
KI	Künstliche Intelligenz.
LOM	Last of Maximum.
MAD	Mittlere Absolute Abweichung (Mean Absolute Deviation).
ML	Maschinelles Lernen (Machine Learning).
MOM	Middle of Maximum.
OHE	One-Hot Kodierung (One-Hot Encoding).
PDP	Partial Dependence Plot.
PFI	Permutation Feature Importance.
SHAP	SHapley Additive exPlanations.
UI	Benutzerschnittstelle (User Interface).
UX	User Experience.
XAI	Erklärbare Künstliche Intelligenz (Explainable Artificial Intelligence).
XAIR	XAI-Empfehlungssystem (XAI Recommender).

# 1. Einleitung

Das maschinelle Lernen (ML), ein Bereich der Künstlichen Intelligenz (KI), der mithilfe verschiedener mathematischer Methoden selbstständig Wissen aus einer großen Menge an Daten extrahiert, gewinnt zunehmend an Bedeutung. Fortschritte in der Computertechnologie, ein rasanter Anstieg verfügbarer Datenmengen, sowie die Generalität der ML-Techniken macht diese für eine Vielzahl von Anwendungen sehr attraktiv. Bereits heute beeinflussen sie sämtliche Branchen und Lebensbereiche. (Hamon et al. 2020)

Dieses automatisierte, maschinelle Lernen komplexer Strukturen und Beziehungen aus sehr großen Datensätzen ist sehr leistungsfähig und gilt als Schlüsseltechnologie für kognitive Systeme (Döbel et al. 2018). In einer Online-Umfrage des Unternehmens McKinsey vom Juni 2020 mit 2390 Teilnehmern aus Unternehmen verschiedener Größen, Branchen und Regionen, gaben 48% an, KI in mindestens einer Funktion eingesetzt zu haben (McKinsey Global Publishing 2020).

Um die gewünschte Genauigkeit der Vorhersagen zu erreichen, werden statt menschenverständlicher Entscheidungssysteme oftmals bspw. tiefe neuronale Netze (DNNs) verwendet. Aufgrund ihrer Komplexität sind diese für den Menschen allerdings oft unverständlich und ihre Entscheidungen nicht nachvollziehbar. Die mangelnde Transparenz dieser sogenannten Black-Boxes ist ein großer Nachteil und ein Hindernis bei ihrem Einsatz. Besonders in kritischen Bereichen wie bei dem autonomen Fahren, im Strafvollzug, im Militär oder in der Medizin ist eine Erklärung bzw. Rechtfertigung der Verhaltensweisen und Entscheidungen für die Nachvollziehbarkeit, Fairness und Sicherheit einer ML-Anwendung essenziell. (Arrieta et al. 2020)

## 1.1. Problemstellung

Erklärbare Künstliche Intelligenz (XAI, Explainable Artificial Intelligence) versucht mithilfe von verschiedenen Methoden, das Problem fehlender Transparenz von ML-Modellen zu adressieren und die Ergebnisse der Lösung für den Menschen verständlich zu machen (Adadi & Berrada 2018). In der jüngsten Forschung wurde eine Vielzahl von XAI-Methoden und -Implementierungen in Form von eigenständigen Prototyp-Lösungen vorgestellt (Vilone & Longo 2020). Jedoch werden diese in der Praxis kaum angewendet. Dies liegt zum einen daran, dass XAI ein neuer, sich rasant entwickelnder Bereich ist (Belle & Papantonis 2020, S. 20), und zum anderen, dass das bereits existierende Wissen verstreut ist und organisiert werden muss (Vilone & Longo 2020, S. 1).

Einige wissenschaftliche Publikationen klassifizieren zwar diverse Methoden, liefern allerdings selten konkrete Richtlinien oder Ratschläge für ihre Anwendung. Staatliche Behörden wie bspw. das Bundesamt für Sicherheit in der Informationstechnik (BSI) weist auf die Relevanz der Erklärbarkeit von KI-Systemen hin. Außerdem gibt das BSI an, dass bei der Auswahl methodenspezifische Eigenschaften der Eingaben berücksichtigt, und plausible Erklärungen ermöglicht werden müssen (Federal Office

for Information Security 2021, S. 41). Es bietet allerdings keine Informationen, die den Nutzer zu der erfolgreichen Auswahl und Anwendung einer in diesem Sinne geeigneten XAI-Methode verhelfen. Nur wenige der für die ethische Nutzung von Algorithmen und KI veröffentlichten Prinzipien, Selbstverpflichtungen und Frameworks enthalten Empfehlungen oder Beispiele für die Operationalisierung dieser Prinzipien (AlgorithmWatch 2019).

Zusammenfassend existieren zwar diverse XAI-Methoden, aber keine konkrete Empfehlung, in welchem Kontext der Einsatz einer oder mehrerer bestimmter sinnvoll ist. Außerdem gibt es keine Übersicht über die Methoden oder bewährte Verfahren für ihre Auswahl oder Implementierung, die den Nutzer zu ihrer tatsächlichen Anwendung befähigen.

## 1.2. Zielsetzung

Im Rahmen dieser Masterarbeit wird zur Förderung des Einsatzes von XAI daher die Frage beantwortet:

*Wie kann man den Nutzer bei der Auswahl und anschließenden Anwendung von geeigneten XAI-Methoden auf Black-Box Modelle durch Empfehlungen unterstützen?*

Um die Auswahl der anzuwendenden XAI-Methoden zu erleichtern, soll der Nutzer diesbezüglich Empfehlungen für Methoden erhalten, die sich für seinen spezifischen Anwendungskontext eignen. Um ihm eine effektive Unterstützung bieten zu können, muss das Einholen einer Empfehlung jederzeit und ohne viel Aufwand möglich sein.

Um die Forschungsfrage zu beantworten, wird im Rahmen dieser Arbeit ein XAI-Empfehlungssystem konzipiert und implementiert. Für ein Klassifikations- oder Regressionsmodell, das mit tabellarischen Daten arbeitet, erhält der Nutzer ohne tiefe ML-Kenntnisse nachvollziehbar begründete Empfehlungen geeigneter XAI-Methoden. Für die Nutzung des Systems ist Domänenwissen und Wissen über die Eigenschaften des Anwendungskontexts notwendig, allerdings werden keine tieferen ML-Kenntnisse vorausgesetzt.

Durch eine Bereitstellung in Form einer Web-Anwendung ist das XAI-Empfehlungssystem stetig verfügbar. Der Nutzer kommt schnell in den Genuss einer begründeten Empfehlung und bleibt von der Komplexität der Informationsbeschaffung und Eignungsbeurteilung der XAI-Methoden verschont. Um die Bereitschaft der tatsächlichen Anwendung der empfohlenen XAI-Methoden zu erhöhen, liegt der Fokus auf der Umsetzbarkeit der Methodenvorschläge durch Empfehlungen technischer und methodischer Aspekte, statt auf einer großen Anzahl im System verfügbarer XAI-Methoden.

## 1.3. Methodik und Aufbau der Arbeit

Nachfolgend wird die zur Beantwortung der Forschungsfrage verwendete Methodik aufgeführt und eine Übersicht über die Inhalte der Kapitel gegeben.

Um dem gestaltenden, explorativen Charakter der Forschungsfrage gerecht zu werden, wird für die Erstellung dieser Arbeit der lösungsorientierte Design Science Research (DSR) Ansatz in Erwägung gezogen, der durch Hevner et al. (2004) für Informationssysteme eingeführt wurde. Auf Basis wissenschaftlicher und praktischer Kriterien werden iterativ und inkrementell Designartefakte erzeugt und evaluiert. Diese können innovative Ideen, Praktiken, technische Fähigkeiten und Produkte sein. Zudem bietet dieses Vorgehen datenwissenschaftliche Forschungsrichtlinien, die vielen methodischen Frameworks als Grundlage dienen. (Hevner et al. 2004)

Da der DSR Ansatz Defizite bei der methodischen Unterstützung der Problemformulierung und -darstellung hat, orientiert sich diese Arbeit an dem Requirements Driven DSR Framework von Braun et al. (2015). Bei diesem wird der DSR Ansatz um eine Anforderungsanalyse erweitert. Die erhobenen Anforderungen werden für die Operationalisierung von Forschungszielen und zur Dokumentation des Forschungsfortschritts verwendet, was sich sehr gut für die Implementierung eines Prototyps eignet. Die Evaluation des Artefakts findet, wie in (Requirements Driven) DSR vorgesehen, ebenfalls iterativ und anhand der Anforderungen statt. (Braun et al. 2015)

Um einen Überblick über den Bereich XAI, XAI-Methoden und Empfehlungssysteme zu erhalten, wird zudem eine Literaturrecherche nach vom Brocke et al. (2009) durchgeführt. Dafür wird mittels einer Stichwortsuche zuerst relevante Literatur mit einem Fokus auf möglichst aktuelle und wissenschaftliche Werke identifiziert. Ausgehend von Molnar (2019), der einen guten Überblick über XAI-relevante Terminologien und populäre XAI-Methoden bietet, wird eine Rückwärts- bzw. Vorwärtssuche angewandt. Bei der Rückwärtssuche werden die Referenzen in dem Artikel analysiert, bei der Vorwärtssuche die Artikel, welche Referenzen auf die untersuchte Arbeit enthalten (vom Brocke et al. 2009).

Die für diese Arbeit erforderlichen Grundlagen sind in Kapitel 2 aufgeführt. Dort wird die Ist-Situation analysiert. Außerdem wird auf andere Arbeiten und existierende kommerzielle Lösungen eingegangen, welche sich mit verwandten Fragestellungen beschäftigen.

In Kapitel 3 werden zunächst die Forschungsfragestellung und die dafür zu klärenden Teilfragen aus den aktuellen Problemen abgeleitet. Außerdem werden Ziele und Annahmen des XAI-Empfehlungssystems formuliert. Um ein Gesamtbild des Problems aus Sicht des Nutzers zu erhalten, wird zudem eine Anforderungsanalyse in Form von qualitativen, halbstrukturierten Online-Einzelinterviews durchgeführt. Diese aufgenommenen Anforderungen werden während der Konzeptionsphase (Kapitel 4) aufgrund des iterativen und inkrementellen DSR-Vorgehens weiter überarbeitet.

In Kapitel 4 wird zunächst geklärt, wie die Beurteilung der Eignung einer XAI-Methode erfolgt. Anschließend wird anhand ausgewählter Methoden ein Voraussetzungs- und Kriterienkatalog zur Eignungsbeurteilung erstellt. Nach Auswahl einer für die identifizierten Kriterien adäquaten Wissensrepräsentation wird näher auf die Konfiguration und den Aufbau des Empfehlungssystems eingegangen.

In Kapitel 5 werden die für die Umsetzung relevanten Technologien genannt. Außerdem bietet es verschiedene Einblicke in die konkrete Implementierung des Prototyps auf Basis des zuvor erläuterten Konzepts.

In Kapitel 6 wird das prototypisch implementierte XAI-Empfehlungssystem qualitativ in halbstrukturierten Online-Einzelinterviews anhand der Erfüllung der in Kapitel 3 definierten Anforderungen und Systemzielen evaluiert. Die Ergebnisse werden anschließend hinsichtlich der Beantwortung der Forschungsfrage, der Einordnung in den Stand der Forschung und ihrer Limitationen und zukünftigen Herausforderungen interpretiert.

Kapitel 7 fasst die Arbeit in einem Fazit zusammen und gibt einen Ausblick auf zukünftige Erweiterungsmöglichkeiten des erstellten XAI-Empfehlungssystems.

## 2. Expertensystem für die Auswahl einer XAI-Methode

Dieses Kapitel befasst sich mit der Beschreibung der Grundlagen, auf denen diese Arbeit aufbaut. Zuerst wird auf die Black-Box Problematik heutiger ML-Verfahren eingegangen und der daraus resultierende Bereich der XAI vorgestellt. Dabei werden grundlegende Begriffe zur taxonomischen Einordnung von XAI-Methoden erläutert.

### 2.1. Machine Learning und seine Black-Box Problematik

Als KIs werden Methoden bezeichnet, die rational und autonom schlussfolgern, handeln oder entscheiden können und/oder sich an komplexe oder verändernde Umgebungen anpassen. Beim ML, einem Teilbereich der KI, werden durch diverse mathematische Techniken der Algorithmik, Statistik und Optimierungstheorie Informationen und Beziehungen aus verschiedensten Datenmengen extrahiert, um ein damit in Verbindung stehendes Problem zu lösen. Aufgaben der Klassifizierung, Erkennung, Generierung etc. können somit anhand verschiedener Datenformate, bspw. Bildern, tabellarischer Daten oder Text, erledigt werden. Ein ML-Modell kann auf diese Weise neue, unbekannte Eingabedaten mithilfe des Wissens über andere, im Trainingsprozess bereits gesehene Dateninstanzen, einordnen/erkennen/generieren. (Hamon et al. 2020, S. 10)

Bei diesen ML-Modellen kann zwischen transparenten White-Boxes und komplexitätsbedingt intransparenten Black-Boxes unterschieden werden (Vilone & Longo 2020). Die Transparenz eines ML-Modells ist dabei abhängig von seiner Simulierbarkeit durch den Menschen, seiner Zerlegbarkeit in erklärbare Teile (z.B. in Eingaben, Parameter und Berechnungen), sowie von seiner algorithmischen Transparenz, bei der die Ergebniserzeugung nachvollzogen werden kann (Belle & Papantonis 2020). Transparente Modelle, wie bspw. die lineare/logistische Regression, Entscheidungsbäume und regelbasierte Systeme, sind durch ihre eingeschränkte Komplexität intrinsisch interpretierbar (Molnar 2019). Interpretierbarkeit ist der Grad, inwieweit ein Beobachter die Ursache einer Entscheidung verstehen kann (Miller 2019).

Black-Box Modelle, wie u.a. Random Forests, Support Vector Machines und DNNs, sind aufgrund ihrer Komplexität nicht interpretierbar und ihre Entscheidungen daher nicht nachvollziehbar. Es existiert Trade-Off zwischen der Leistung des Modells und seiner Interpretierbarkeit (Arrieta et al. 2020): Auf Kosten der Interpretierbarkeit erreichen diese eine höhere Genauigkeit der Vorhersagen durch Finden komplexerer Entscheidungsgrenzen (Belle & Papantonis 2020). Da sie außerdem komplexere Eingabeparameter wie Bild- oder Textdaten erlauben und einen effizienteren Lernalgorithmus besitzen, werden sie immer häufiger, auch in kritischen Kontexten, eingesetzt (Arrieta et al. 2020).

Viele Beispiele zeigen immer wieder die Unvollkommenheit bereits eingesetzter KI-Systeme auf. Diese

reichen von geschlechterspezifischen Stereotypen bei der Verarbeitung natürlicher Sprache (Bolukbasi et al. 2016), über die Benachteiligung von Frauen beim automatisierten Einstellungsprozess bei Amazon (Dastin 2018) bis hin zu Rassismus bei einem in den USA eingesetzten Algorithmus, der das Strafmaß durch die Vorhersage der Wahrscheinlichkeit einer erneuten Straftat bestimmt (Angwin et al. 2016).

Solche u.a. diskriminierende Fehlentscheidungen können bspw. durch Verzerrungen (Bias) verursacht werden, die in vielen verschiedenen Formen auftreten können, vgl. Mehrabi et al. (2019).

Bei der Datenerhebung können sie bspw. anfallen, wenn die Trainingsdaten nicht zufällige Stichproben sind (Stichprobenverzerrung) oder sie nicht die Realität repräsentieren (Repräsentationsverzerrung). Auch bei einer perfekten Stichprobenziehung und Auswahl der Eingabevariablen (Features) können die Daten durch historische und sozio-technische Probleme, z.B. aufgrund sensibler Features wie ethnische Herkunft oder Geschlecht, verzerrt sein (historische Verzerrung). Modellbezogene Verzerrungen können auftreten, wenn bspw. durch falsche Annahmen über die Daten einer Personengruppe die Definition des Modells und das Vorhersageergebnis einer individuellen Person beeinflusst wird (Verzerrung durch Aggregation). Sie können auch unabhängig der Trainingsdaten durch den Algorithmus selbst entstehen oder bspw. durch die Verwendung von unangemessenen und unverhältnismäßigen Evaluationsmetriken. (Mehrabi et al. 2019)

Im Internet sind zahlreiche Sammlungen dokumentierter Fehler von in der realen Welt eingesetzten KI-Systemen zu finden; die bekannte AI Incident Database enthält bereits über 1200 Einträge riskanter Zwischenfälle (Partnership on AI 2021).

Auch Regierungen reagieren auf den Bedarf an Transparenz dieser Systeme: Im Mai 2018 trat die europaweit geltende Datenschutzgrundverordnung (DSGVO) in Kraft, welche bei automatisierten Entscheidungsfindungen verlangt, dass Betroffenen „aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung“ (Art. 13 Abs. 2 lit. f, Art. 14 Abs. 2 lit. g) verfügbar gemacht werden müssen. Laut Artikel 22 hat der Betroffene das Recht, nicht ausschließlich einer, auf automatisierter Verarbeitung beruhenden Entscheidung mit beeinträchtigender Wirkung unterworfen zu werden (Art. 22 Abs 1). Zumindest hat er jedoch das Recht auf menschliche Intervention, auf Darlegung des eigenen Standpunkts und auf eine Anfechtung der Entscheidung (Art. 22 Abs. 4). (Europäisches Parlament und Rat der Europäischen Union 2016).

Im Hinblick auf bestehende Probleme und einen vermehrten, produktiven Einsatz von ML-Modellen, muss das Problem mangelnder Transparenz und Nachvollziehbarkeit dringend adressiert werden. Eine Erklärung bzw. Rechtfertigung der Verhaltensweisen und Entscheidungen ist für die Nachvollziehbarkeit, Fairness und Sicherheit der Modelle essenziell (Arrieta et al. 2020).

## 2.2. Explainable Artificial Intelligence (XAI)

Das Forschungsfeld der XAI versucht die Interpretierbarkeit von und das Vertrauen in ML-Modelle zu fördern, ohne ihre (Lern-)Leistung einzuschränken (Arrieta et al. 2020). Dafür werden Methoden und Techniken entwickelt, die Aspekte der Datenwissenschaft, des MLs, der Sozialwissenschaft und

der Mensch-Computer-Interaktion (HCI, Human-Computer Interaction) berücksichtigen. Der Begriff „XAI“ trat erstmals 2004 auf, obwohl das Problem der Interpretierbarkeit von KI-Systemen seit der Anwendung der ersten Expertensysteme in den 1970ern existiert. Durch den zunehmenden Einsatz von ML in allen Branchen und ihren wesentlichen Einfluss in kritischen Entscheidungsprozessen sind die Forschungsanstrengungen im Bereich XAI in den letzten Jahren gestiegen. (Adadi & Berrada 2018)

Je komplexer und dadurch unverständlicher ein ML-Modell wird und je wichtiger seine vom Anwendungskontext abhängende Fehlerresistenz ist, desto wichtiger ist seine Interpretierbarkeit. Sie ist essenziell, wenn Menschen die KI-generierten Ergebnisse verstehen, ihnen angemessen vertrauen und sie verwenden sollen. (Adadi & Berrada 2018)

Erklärungen der ML-Modelle sind sinnvoll, um neue Erkenntnisse der Problemdomäne zu erhalten. Zudem sind sie wichtig bzw. teilweise gesetzlich vorgeschrieben, damit Entscheidungen gerechtfertigt werden können. Außerdem helfen sie dabei, das Systemverhalten zu verstehen, Verzerrungen in den Daten zu identifizieren und das Modell dadurch zu verbessern und robuster gegenüber Schwachstellen und Angriffen zu machen. (Arrieta et al. 2020)

Im vorigen Kapitel wurde bereits anhand der Modellkomplexität zwischen transparenten und intransparenten Modellen unterschieden. Transparente Modelle sind intrinsisch interpretierbar und können als „Explainable AI“ verstanden werden. Da sich diese Arbeit auf Empfehlungen für die Anwendung von XAI-Methoden auf existierende (Black-Box) ML-Modelle beliebiger Komplexität bezieht, liegt der Fokus auf XAI in Form von Methoden, die nachträglich, „post-hoc“ angewendet werden. Nachfolgend werden diese Methoden hinsichtlich mehrerer Eigenschaften taxonomisch eingeordnet.

### 2.2.1. Taxonomie interpretierbarer Methoden

Auf intransparente Modelle anwendbare XAI-Methoden unterscheiden sich hinsichtlich der Modellart, auf die sie angewendet werden können, und anhand des Umfangs und des Formats der resultierenden Erklärung:

Es gibt *modellspezifische* und *modellagnostische* XAI-Methoden. Modellspezifische sind auf eine bestimmte Modellart beschränkt, während modellagnostische post-hoc auf jeden Modelltypen angewendet werden können. Da Modellagnostische keinen Zugriff auf die Modellinterna haben, behandeln sie es als Black-Box und liefern die Erklärungen nur anhand einer Analyse der Ein- und Ausgaben. (Molnar 2019)

XAI-Methoden können außerdem nach dem Umfang der erhaltenen Erklärung klassifiziert werden, wobei zwischen *globalen* und *lokalen* Methoden unterschieden wird.

Lokale Methoden erklären die Gründe einer Modellentscheidung für eine einzelne Dateninstanz. Die Gründe sind spezifisch für diese, weshalb die resultierende Erklärung nicht auf eine globale Skala generalisiert werden kann. Die Methode approximiert dafür das Verhalten eines Modells um die beobachtete Dateninstanz herum und erklärt, wie es beim Auftreten solcher Instanzen reagiert. (Belle & Papantonis 2020)



Globale Erklärungen zielen auf ein Verständnis des Verhaltens des Gesamtsystems ab. Eine holistische Interpretierbarkeit ist allerdings schwer zu erreichen, da eine solche Erklärung aufgrund ihrer Komplexität für den Menschen schwer greifbar wäre. Daher gibt es globale Erklärungen auf modularer Ebene, welche sich auf erklärbare Teile, bspw. die Eingaben oder Parameter (z.B. Gewichte im Fall einer linearen Regression) konzentrieren. (Molnar 2019)

XAI-Methoden lassen sich zudem bzgl. ihrer Erklärungsansätze grob in vier Kategorien einteilen, welche nachfolgend kurz erläutert werden:

- *Visuelle Erklärungen* vereinfachen durch Visualisierungen das Verständnis eines Modells und eignen sich dadurch auch für Menschen ohne Expertenwissen. Sie helfen, Einblicke bzgl. der Entscheidungsgrenzen oder der Interaktionen der Features zu erhalten. (Belle & Papantonis 2020)
- Bei einer *Erklärung durch Vereinfachung* basieren fast alle Techniken auf Regelextraktion. Die Regeln werden unter Verwendung der Ein- und Ausgaben konstruiert, die den Entscheidungsprozess approximieren und ihn somit erklären. Bei der anderen Technik der Modellvereinfachung wird ein Black-Box Modell mit einem transparenten, interpretierbaren Modell approximiert. (Adadi & Berrada 2018)
- Eine *Erklärung durch Feature Relevanz* quantifiziert den Einfluss jedes Eingabe-Features auf die Modellvorhersage, indem die Auswirkung der Änderung seines jeweiligen Werts auf das Vorhersageergebnis oder die Modell-Leistung (z.B. die Vorhersagegenauigkeit) beobachtet wird (Adadi & Berrada 2018).
- Für eine *beispielbasierte Erklärung* werden, für den Erhalt eines besseren Verständnisses des Modells, repräsentative Dateninstanzen des Trainingsdatensatzes ausgewählt. Diese zeigen die inneren Beziehungen und Korrelationen des Modells auf. (Arrieta et al. 2020)

### 2.2.2. Ist-Situation und verwandte Arbeiten

Nachfolgend werden die in einer Literaturrecherche ermittelten und im Rahmen der Fragestellung relevanten Arbeiten aufgeführt, welche die Auswahl und Anwendung geeigneter XAI-Methoden auf Black-Box Modelle erleichtern. Dabei wird berücksichtigt, dass die Arbeiten Auskunft über model-lagnostische und daten- und domänenunabhängige Verfahren geben, und dass die Zielgruppe der XAI Empfehlungen keine tieferen ML-Kenntnisse aufweist.

Betrachtet werden dafür wissenschaftlichen Zusammenfassungen des Forschungsbereichs XAI und Richtlinien zur Förderung von Interpretierbarkeit in KI-Systemen. Zudem werden kommerzielle Tools begutachtet, die dem Nutzer die manuelle Anwendung von XAI abnehmen.

#### **Wissenschaftliche XAI Zusammenfassungen – wie zugänglich ist der Wissensstand von XAI für den Nutzer?**

Der Forschungsbereich XAI hat in den letzten Jahren beträchtliche Aufmerksamkeit erlangt und es wurden zahlreiche, eigenständige XAI-Methoden vorgestellt. Wissenschaftliche Literaturüberblicke

versuchen, das sehr verstreut liegende Wissen des sich schnell entwickelnden Bereichs XAI zu bündeln.

Hinsichtlich der Anwendung (geeigneter) spezifischer XAI-Methoden geht Molnar (2019) auf Umstände ein, die das Methodenergebnis negativ oder verfälschend beeinflussen können. Allerdings erhält der Nutzer aufgrund einer fehlenden Zusammenfassung dieser Umstände und ihrer Konsequenzen keine direkte Unterstützung bei der Selektion einer XAI-Methode, die sich für seinen Kontext eignet. Auch Hall & Gill (2019) beschreiben geläufige Interpretierungstechniken tabellarisch und bieten Vorschläge, in welchem Szenario sich ihre Anwendung eignet. Diese Szenarien sind allerdings abstrakt und unvollständig beschrieben; viele der in den originalen Methoden-Veröffentlichungen oder durch Molnar (2019) aufgeführten Anwendungslimitierungen sind nicht erwähnt. Dadurch besteht für den Nutzer bei der Auswahl einer geeigneten XAI-Methode weiterhin ein erhöhter Aufwand bei der Beschaffung der dafür relevanten Informationen. Er erhält keinen konkreten Aufschluss darüber, welche kontextspezifischen Umstände es dabei zu beachten gilt und muss sich intensiv mit den verschiedenen Verfahren auseinandersetzen. Meist ist dafür eine Vertrautheit mit ML-Algorithmen vorausgesetzt.

### **Veröffentlichte Richtlinien oder die Regulatorik – bieten sie Unterstützung bei der Auswahl von XAI-Methoden?**

Viele von Behörden und Unternehmen erstellte Richtlinien und Prinzipien zur Förderung der Transparenz, Gleichheit, Rechenschaftspflicht und Sicherheit von KI-Systemen weisen auf die Relevanz ihrer Interpretierbarkeit hin. Nur wenige bieten allerdings Empfehlungen oder Beispiele, wie diese Prinzipien operationalisierbar sind (AlgorithmWatch 2019). Laut BSI sollen spezifische Eigenschaften der Eingaben für die Auswahl einer XAI-Methode berücksichtigt, und plausible Erklärungen ermöglicht werden (Federal Office for Information Security 2021, S. 41). Allerdings bleiben diese genannten, zu beachtenden Eigenschaften unklar.

Auch andere Richtlinien schlagen methodische Schritte zur Implementierung von Interpretierbarkeit in KI-Systemen vor. Sie konzentrieren sich jedoch hauptsächlich auf das Gesamtbild der Förderung der Interpretierbarkeit und bieten dem Nutzer, als Anwender einzelner XAI-Methoden, wenig Hilfe bei der Durchführung. Leslie (2019) bspw. rät unkonkret zur Anwendung von Methoden, die dem Nutzer dabei helfen, besser informierte und evidenzbasierte Urteile zu fällen und die das Verhalten des Modells plausibel und vernünftig darstellen. Eine umsetzbare Empfehlung wird aber nur bzgl. des Erklärungsumfangs gegeben: Dabei sollte „local-first“ gedacht werden (Leslie 2019, S. 54). Kangur (2020) empfiehlt eine Erklärung durch eine globale Feature Relevanz Methode und die Anwendung einer lokalen Methode auf richtig und falsch klassifizierte Dateninstanzen; Ratschläge bzgl. konkreter Methoden bleiben allerdings auch hier aus. Obwohl auch hier zur Anwendung mehrerer Methoden geraten wird, gibt es aufgrund der Segmentierung keine Orientierung, wie man die dedizierten Methoden für den Erhalt einer „kompletteren“ Erklärung kombinieren kann (Belle & Papantonis 2020).

In der im April 2021 veröffentlichten Studie von Kraus et al. (2021) wird erstmals eine Orientierungshilfe zur Auswahl von Erklärungsstrategien in Form eines Orientierungsbaumes vorgestellt. Dieser gibt in den Blättern konkrete Methodenempfehlungen anhand von in Knoten aufgeführten Eigenschaften, welche insbesondere die Zielgruppen der Erklärung und den Daten- und Modelltypen umfassen. Eine

Rolle bei der Auswahl spielen außerdem die Erfahrung des Anwenders mit der Verwendung von XAI und die Präferenzen, ob die XAI-Methode ein industrieller Standard sein soll, ob sie eine hohe Leistung haben soll oder ob beispielbasierte Erklärungen gewünscht sind. (Kraus et al. 2021, S. 70)

Viele Blätter des Orientierungsbaumes enthalten entweder einzelne, in mehreren Blättern genannte XAI-Methoden, oder mehrere, die nicht ihrer Eignung nach sortiert sind. In letzterem Fall wird die Menge geeigneter Methoden zwar eingegrenzt, allerdings müssen aus diesen wiederum mit erhöhtem Rechercheaufwand Passende ausgewählt werden. Wie die Autoren bereits darlegen, „berücksichtigt [der Orientierungsbaum] zwar die meistzitierten Ansätze (im Falle ihrer praktischen Anwendbarkeit), stellt damit jedoch lediglich eine Momentaufnahme für den Stand der Technik dar“ (Kraus et al. 2021, S. 84). Er ist statisch und droht unübersichtlich zu werden, sollten neue XAI-Verfahren oder eignungsbeeinflussende Eigenschaften als Knoten aufgenommen werden. Obwohl innerhalb der Umfrage ihrer Studie das Problem identifiziert wird, dass Anwender die Funktionsweisen von post-hoc Methoden nicht intuitiv verstehen und dadurch die korrekte Erklärungsinterpretation nicht gesichert ist (Kraus et al. 2021, S. 68), wird dieses nicht adressiert. Die Empfehlung erfolgt ohne Angabe einer Begründung oder von weiteren Methodeninformationen, die zur Anwendung der vorgeschlagenen Erklärungsstrategie befähigen.

### **Kommerzielle Tools – kann man XAI-Orientierung kaufen?**

Es gibt bereits einige kommerzielle Tools, die den Nutzer bei der Anwendung von XAI-Methoden unterstützen (Google 2020b, H2O.ai 2020, DataRobot 2020, Uppington 2020, Spinner et al. 2020). Mit diesen kann XAI automatisiert, teilweise innerhalb einer ML-Pipeline, angewendet werden. In solchen ML-Pipelines werden die separaten, modularen Schritte der Datenaufbereitung, des Trainings, der Evaluierung und der Veröffentlichung des Modells automatisiert sequenziell ausgeführt. Bei der Nutzung eines solchen kommerziellen Tools ist der Nutzer allerdings an die jeweilige Plattform gebunden und auf die Anwendung weniger vorimplementierter XAI-Methoden beschränkt. Diese Methoden übermitteln ihre Ergebnisse meist in komplexen Visualisierungen, welche ML- und datenwissenschaftliche Fachkenntnisse voraussetzen. Der Nutzer erhält keine Empfehlung, welche Methoden sich aufgrund des Modell-, Daten- und Nutzungskontexts am besten eignen, da diese vor der automatisierten Anwendung nicht berücksichtigt werden.

Insgesamt schaffen wissenschaftliche Literaturzusammenfassungen einen guten Überblick existierender XAI-Methoden, bieten allerdings keine Orientierung bei der Auswahl einer passenden. Das Lesen und Verstehen setzt zudem ML- und datenwissenschaftliche Fachkenntnisse voraus.

Ein existierender Orientierungsbaum bietet zwar Unterstützung bei der ersten Auseinandersetzung mit XAI-Methoden, schließt aber lediglich unbegründet Methoden aus der Empfehlungsmenge aus und bietet keine Hilfe bei ihrer Anwendung.

Obwohl Behörden und Unternehmen auf interpretierbare ML-Modelle drängen, geben sie auch keine umsetzbaren Richtlinien oder konkrete Vorschläge zur Verwirklichung vor. Die Frage, unter welchen spezifischen Umständen des Modell-, Daten- und Nutzungskontexts sich welche XAI-Methode wie gut für eine Anwendung eignet, kann einem Nutzer keine bisher veröffentlichte, schriftliche Arbeit ohne erheblichen Rechercheaufwand und Fachwissen beantworten. Weder wurden eignungsbeeinflussende Kriterien ermittelt anhand deren man die XAI-Methoden konkret vergleichen kann, noch fand eine

Beurteilung ihres Einflusses auf die Methodeneignung statt.

Außerdem gibt es kein existierendes System, das dieses Wissen für die Anwendung einer XAI-Methode berücksichtigt. Kommerzielle Tools nehmen dem Nutzer durch Automatisierung zwar viel Arbeit bei der Anwendung von XAI-Verfahren ab, limitieren ihn jedoch auf die Nutzung dieser speziellen Plattform und auf die Verwendung der darin implementierten, aufgrund des spezifischen Nutzungskontexts eventuell ungeeigneten Methoden.

In dieser Arbeit wird der Ansatz verfolgt, die aktuellen, oben genannten Herausforderungen durch die Implementierung eines Empfehlungssystems zu lösen. Nachfolgend werden daher die dafür notwendigen Grundlagen vorgestellt.

## 2.3. Empfehlungssysteme und Expertensysteme

Heute sind Empfehlungssysteme eine zentrale Komponente vieler Online-Shops, um den Kunden mit geringem Interaktionsaufwand zu Produkten zu führen, die er aufgrund seiner Interessen und Bedürfnisse mit hoher Wahrscheinlichkeit kauft. Ihr allgemeines Ziel ist es, für den Nutzer aus einer großen, unüberschaubaren Menge von Objekten automatisiert Alternativen auszuwählen und ihm vorzuschlagen. Zur Steigerung der Nutzerzufriedenheit berücksichtigen sie nicht nur seine individuellen Präferenzen, sondern auch Kontextinformationen, bspw. seine situationsabhängigen Ziele. (Ziegler & Loepp 2019)

Für die Empfehlungsgenerierung gibt es drei Verfahren. Beim kollaborativen Filtern wird das gemeinsame Feedback anderer, ähnlicher Nutzer für eine passende Empfehlung genutzt. Bei inhaltsbasierten Methoden wird der Inhalt der Objekte bzw. deren Eigenschaften analysiert. Dem Nutzer werden Objekte vorgeschlagen, die ähnlich denen sind, mit denen er bereits interagiert hat. Wissensbasierte Methoden greifen auf tiefergehendes Domänenwissen hinsichtlich der Objekte zurück. Sie liefern Vorschläge, die den Anforderungen des Nutzers an das Objekt genügen, weshalb semantisch tiefergehende Erklärungen für die Empfehlung bestimmter Objekte produziert werden können. Für die Erstellung wissensbasierter Systeme ist das Zusammentragen von detailliertem (Experten-)Wissen über Charakteristika der Objekte notwendig. (Ziegler & Loepp 2019)

Da bei dem kollaborativen Filtern und bei inhaltsbasierten Methoden bereits vergangenes Nutzerfeedback bzw. Empfehlungen berücksichtigt werden, wird im Folgenden näher auf den wissensbasierten Empfehlungsansatz in Form von Expertensystemen eingegangen.

Ein computerbasiertes Expertensystem gehört zu den wissensbasierten Systemen, dient als schnell verfügbarer Ersatz eines Experten mit Fachwissen und kann zeitnah, einfach und kostengünstig für die Lösung von fachlichen Problemen konsultiert werden. Es ist in der Lage, Beratungsanfragen einer spezifischen Problemdomäne entgegenzunehmen, mit Fachwissen begründete Empfehlungen zu erstellen und diese zu kommunizieren und zu begründen. (Holsapple & Whinston 2013)

Ein Expertensystem besteht aus drei Komponenten: Der Wissensbasis, der Inferenzkomponente und

der Interviewerkomponente.

In der Wissensbasis wird das erworbene, für die Problemlösung benötigte Wissen formalisiert und repräsentiert. Man kann dabei zwischen deskriptivem, d.h. Faktenwissen, und schlussfolgerndem Wissen unterscheiden. Das Wissen kann in verschiedenen Repräsentationen zur Verfügung stehen, wobei das schlussfolgernde häufig in Form von Regeln vorliegt. (Holsapple & Whinston 2013)

Die Inferenzkomponente leitet die Schlussfolgerungen entsprechend der Eingaben durch Anwendung der in der Wissensbasis abgelegten, fachspezifischen Regeln und Fakten ab. Sie empfängt die für die Problemlösung benötigten Eigenschaften von der Interviewerkomponente, der Nutzerschnittstelle. Die Schnittstelle zum Nutzer kann funktional außerdem in die Erklärungskomponente, die die Ergebnispräsentation und -begründung übernimmt, und die Wissenserwerbskomponente unterteilt werden. Letztere ermöglicht die Eingabe und Veränderung des Wissens der Wissensbasis entweder direkt durch einen Experten oder durch einen Wissensingenieur, der die Formalisierung übernimmt. (Puppe 1991, S. 12)

Es existieren verschiedene Variationen von Expertensystemen bzgl. ihrer Wissensrepräsentation und -verarbeitung, der (Un-)Sicherheit beim Schlussfolgern oder ihrer Lernfähigkeit. Auch nicht-diskretes Wissen kann in Expertensystemen Anwendung finden. (Holsapple & Whinston 2013) Der Aufbau und die Funktionsweise eines solchen, vages Wissen verwendenden Fuzzy-Expertensystems wird nun genauer erläutert.

## 2.4. Fuzzy-Expertensysteme

Fuzzy-Expertensysteme werden als Erweiterung des klassischen Expertensystems verwendet, um „das Wissen und die Lösungsstrategien von Experten möglichst inhaltserhaltend formal repräsentieren und [...] verarbeiten zu können“ (Nissen 2007, S. 14–15). Durch Verwendung der Fuzzy-Logik, die eine Unschärfe und Nicht-Eindeutigkeit der Schlussfolgerungen von unscharfen Prämissen erlaubt (Zadeh 1975, S. 407), liefern sie ungefähre Ausgaben bzw. Entscheidungen basierend auf vagen, linguistischen Eingaben.

Um dies besser zu verstehen wird zunächst ein grundlegendes Verständnis der Fuzzy-Logik geschaffen, bevor der Aufbau eines solchen Expertensystems genauer erläutert wird.

### 2.4.1. Fuzzy-Logik

Die Fuzzy-Logik, die 1965 von Prof. L. A. Zadeh als Logik des approximativen Schließens erfunden wurde, bietet eine Möglichkeit der mathematischen Modellierung von Unschärfe. Neben der Verwendung in wissensbasierten Anwendungen in Form von Fuzzy-Expertensystemen, ist sie heutzutage in technisch-industriellen Anwendungsbereichen gut etabliert und als „Fuzzy Control“ bei regelungstechnischen Anwendungen durch die Arbeit von Mamdani & Assilian (1975) weit verbreitet.

Um die Fuzzy-Logik zu verstehen gilt es zunächst den Begriff der Fuzzy-Menge zu erläutern. Eine klassische Menge  $A$  ist eine Menge an Objekten (Elementen) mit scharfen (crisp), eindeutigen Grenzen, die innerhalb einer Grundmenge  $X(x \in X)$ , dem Diskursuniversums der Elemente, liegt. Ein Element kann entweder zu einer Menge gehören oder nicht. Die Mitgliedsgrad- bzw. Zugehörigkeitsfunktion

$\mu_A$  einer klassischen Menge  $A$  kategorisiert ein Element  $x$  daher binär in „Vollmitglied“ oder „Nicht-Mitglied“ (Bai & Roth 2019):

$$\mu_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases} \quad (1)$$

Eine unscharfe (fuzzy) Menge  $\tilde{A}$  hat im Gegensatz zu einer klassischen Menge weiche Grenzen, was eine graduelle Zugehörigkeit eines Elements zu dieser im Bereich  $[0,1]$  erlaubt (Böhme 1993, S. 5).

$$\mu_{\tilde{A}}(x) \in [0, 1] \quad (2)$$

Jedem Element der Grundmenge  $X$  wird ein Mitgliedsgradwert für die Fuzzy-Menge  $\tilde{A}$  zugeordnet, wobei nicht zur Menge gehörende den Wert 0 erhalten (Böhme 1993, S. 5). Da alle Elemente von  $X$  mit einer gewissen Zugehörigkeit in der Fuzzy-Menge vorhanden sind, wird diese nicht durch die in ihr enthaltenen Elemente, sondern über die Grade der Elementzugehörigkeiten definiert. Die Darstellung erfolgt als Liste mit zwei Tupeln, dem Element und seinem Zugehörigkeitswert zu  $\tilde{A}$ :

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x)) \mid x \in X\} \quad (3)$$

Klassische Mengen können also als Fuzzy-Menge verstanden werden, bei denen alle Elemente den Mitgliedsgradwert 1 haben (Böhme 1993, S. 5).

Der Unterschied zwischen scharfen und fuzzy Mengen lässt sich am Beispiel der Temperatur gut veranschaulichen. Die Einteilung dieser in die drei Kategorien KALT, WARM und HEISS für das Diskursuniversum  $X = [0; 45]$  sind für die beiden genannten Mengenlehren in Abbildung 2.1 dargestellt.

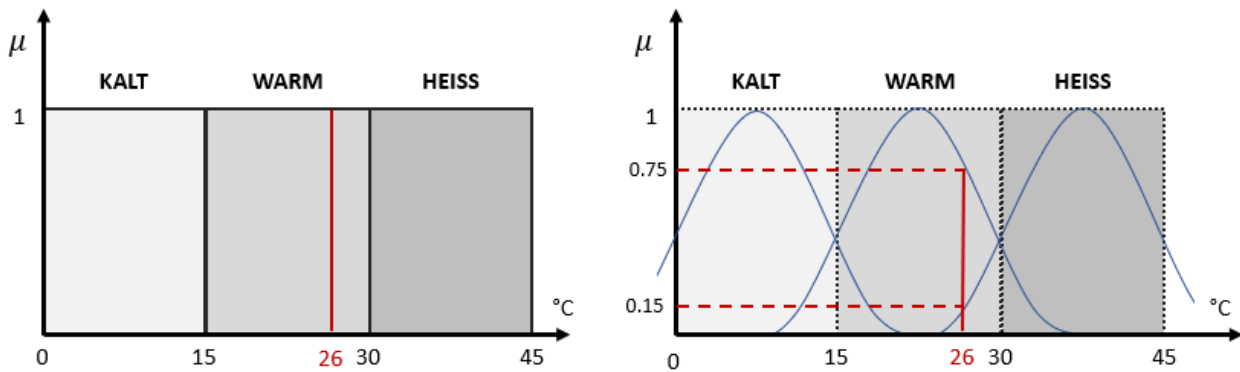


Abbildung 2.1.: Klassische und Fuzzy-Mengen, angelehnt an (Bai & Roth 2019, S. 443)

Während die Einordnungsgrenzen der klassischen Mengenlehre links in der Abbildung 2.1 hart sind und das rot gekennzeichnete Element  $x = 26$  vollkommen der Kategorie WARM zugeordnet wird, erhält es für die Fuzzy-Mengen folgende Mitgliedsgrade:

$$\begin{aligned} \mu_{KALT}(26) &= 0 \\ \mu_{WARM}(26) &= 0.75 \\ \mu_{HEISS}(26) &= 0.15 \end{aligned}$$

Für Abbildung 2.1 wurden beispielhaft gaußförmige Zugehörigkeitsfunktionen definiert, welche sich gut für Systeme eignen, die eine hohe Regelgenauigkeit verlangen. In der Praxis werden häufig dreieckige und trapezförmige Funktionen verwendet. Sie werden unter anderem in Systemen genutzt, die keine sehr hohe Regelgenauigkeit benötigen. (Bai & Roth 2019, S. 446)

Relevante Terminologien im Kontext solcher Zugehörigkeitsfunktionen und ausgewählte Operationen dieser sind in Abbildung 2.2 dargestellt. Die Menge aller Elemente des Diskursuniversums  $X$ , deren Zugehörigkeitsgrad größer als 0 ist, wird als ihre stützende Menge („Support“) bezeichnet (Nissen 2007, S. 9). Ihr Kern („Core“) besteht aus der Menge aller Elemente mit dem Zugehörigkeitsgrad 1 und ihre Grenzbereiche („Boundaries“) aus denen mit einem zwischen 0 und 1 (Bai & Roth 2019, S. 446):

$$\begin{aligned} A_{Support}(\tilde{A}) &= \{x \in X | \mu_{\tilde{A}}(x) > 0\} \\ A_{Core}(\tilde{A}) &= \{x \in X | \mu_{\tilde{A}}(x) = 1\} \\ A_{Boundary}(\tilde{A}) &= \{x \in X | 0 < \mu_{\tilde{A}}(x) < 1\} \end{aligned}$$

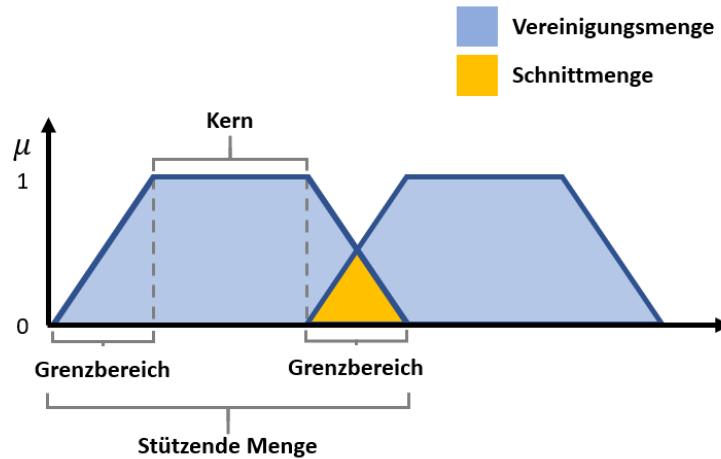


Abbildung 2.2.: Eigenschaften und Operationen von Fuzzy-Mengen

Die grundlegenden Operationen der klassischen Mengentheorie, z.B. die Bildung von Schnittmengen und Vereinigungen, sind wie in Abbildung 2.2 dargestellt, auf Fuzzy-Mengen übertragbar. Demnach werden die Zugehörigkeitsfunktionen dieser für die Fuzzy-Mengen  $\tilde{A}$  und  $\tilde{B}$  folgendermaßen definiert (Böhme 1993, S. 34):

Für die Vereinigungsmenge  $\tilde{A} \cup \tilde{B}$ :

$$\mu_{\tilde{A} \cup \tilde{B}}(x) = \mu_{\tilde{A}}(x) \cup \mu_{\tilde{B}}(x) = \max(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)) \quad (4)$$

Für die Schnittmenge  $\tilde{A} \cap \tilde{B}$ :

$$\mu_{\tilde{A} \cap \tilde{B}}(x) = \mu_{\tilde{A}}(x) \cap \mu_{\tilde{B}}(x) = \min(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)) \quad (5)$$

Bei der Fuzzy-Logik nach Zadeh (1975) entsprechen solche Zugehörigkeitswerte mehrwertigen, linguis-

tischen Wahrheitswerten. Das ermöglicht die Bewertung von Aussagen mit bspw. „wahr, falsch, nicht wahr, sehr wahr, ziemlich wahr, nicht sehr wahr“ und „nicht sehr falsch“ statt wie in der klassischen, zweiwertigen Logik nur mit „wahr“ oder „falsch“ (Zadeh 1975, S. 407).

Fuzzy-logische Junktoren/Operatoren werden dabei folgendermaßen formuliert (Zadeh 1975, S. 410):

$$\mu_{\tilde{A}}(x) \vee \mu_{\tilde{B}}(x) = \mu_{\tilde{A} \cup \tilde{B}}(x) = \max(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)) \quad (6)$$

$$\mu_{\tilde{A}}(x) \wedge \mu_{\tilde{B}}(x) = \mu_{\tilde{A} \cap \tilde{B}}(x) = \min(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)) \quad (7)$$

Das Beispiel der Temperatur in Abbildung 2.1 verdeutlicht, dass die Fuzzy-Logik die Möglichkeit der (mehrfachen) Zuordnung von Objekten zu natürlichsprachigen Ausdrücken mit inhärenter Unschärfe ermöglicht und somit als Versuch gesehen werden kann, die allgegenwärtige Realität der Unschärfe und Vagheit der menschlichen Wahrnehmung mathematisch zu modellieren (Zadeh 1975).

### 2.4.2. Anwendung der Fuzzy-Logik in einem Expertensystem

Für eine praktische Implementierung der Fuzzy-Logik innerhalb eines Expertensystems ist die Ausführung der nachfolgend aufgeführten Schritte notwendig, die schematisch in Abbildung 2.3 dargestellt sind.

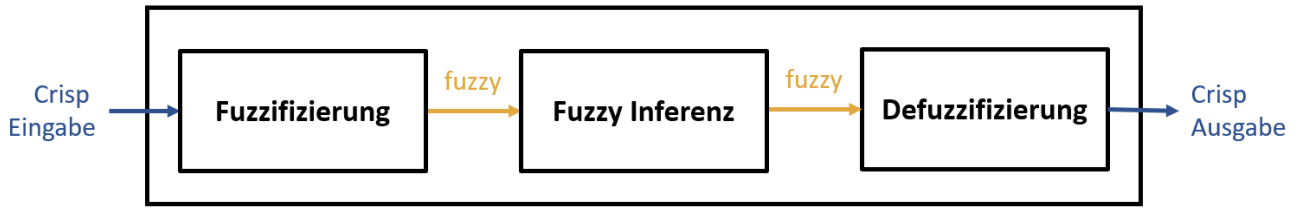


Abbildung 2.3.: Schematischer Aufbau eines Fuzzy-Expertensystems

#### 1. Fuzzifizierung

Bei der Fuzzifizierung werden klassische (crisp) Daten zu fuzzy Daten umgewandelt. Nach der Ermittlung der Eingabe- und Ausgabevariablen des Fuzzy-Systems ist eine Definition von Zugehörigkeitsfunktionen durch die Aufteilung des Diskursuniversums der Variablen in Fuzzy-Mengen notwendig. Diese werden durch linguistische Terme (vgl. Terme KALT, WARM, HEISS der Abbildung 2.1) repräsentiert. (Bai & Roth 2019, S. 446)

#### 2. Fuzzy Inferenz

Das schlussfolgernde Wissen eines Fuzzy-Expertensystems liegt in Form von fuzzy „Wenn-Dann“-Regeln vor. Diese verbinden die Zugehörigkeitsfunktionen der Eingabeparameter in ihrer Prämisse mit denen der Ausgangsvariablen in der Konklusion. Eine gängige Inferenzmethode ist die Max-Min-Inferenz nach Mamdani & Assilian. Dabei werden die Regeln abhängig von der aggregierten Zugehörigkeitshöhe der Eingaben, d.h. vom Erfülltheitsgrad der Gesamtprämisse, aktiviert. Dadurch kann der Wahrheitsgehalt des Ergebnisses nicht höher als der Erfülltheitsgrad der Prämissen sein.

Zur Ermittlung der Erfülltheit der Gesamtprämisse werden die logischen Junktoren der Gleichungen



6 und 7 verwendet. Eine gleichzeitige Aktivierung mehrerer Regeln ist möglich. Sie werden durch ein fuzzy-logisches „oder“ verknüpft und so zu einer Fuzzy-Ergebnismenge zusammengeführt. (Mamdani & Assilian 1975)

Die Max-Min-Inferenz ist am Beispiel der folgenden zwei Regeln in Abbildung 2.4 dargestellt:

Regel 1: Wenn  $X$  „A1“ ist und  $Y$  „B1“ ist, dann ist  $Z$  „C1“

Regel 2: Wenn  $X$  „A2“ ist und  $Y$  „B2“ ist, dann ist  $Z$  „C2“

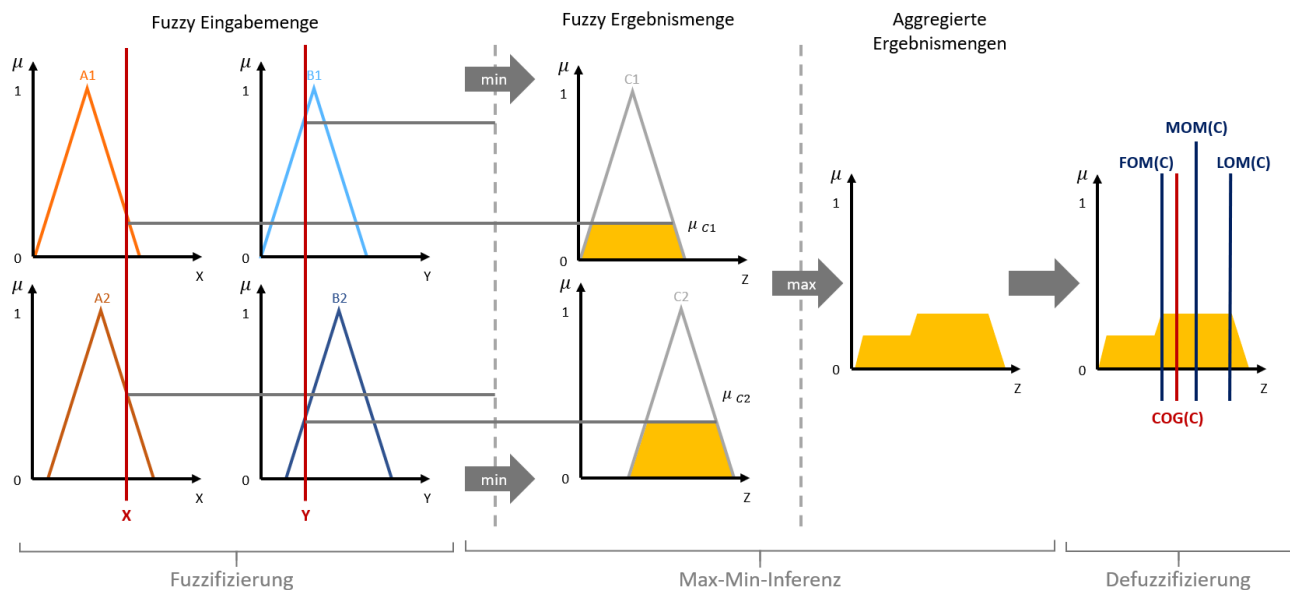


Abbildung 2.4.: Max-Min-Inferenz und Defuzzifizierung, angelehnt an (Cho et al. 2017)

### 3. Defuzzifizierung

Für die Ausgabe muss im letzten Schritt, der Defuzzifizierung, für die Fuzzy-Ergebnismenge der aktivierte Implikationen ein crisp Wert ermittelt werden. Dafür gibt es verschiedene Techniken.

Bei Maxima Methoden, die hauptsächlich in wissensbasierten Fuzzy-Systemen verwendet werden, wird ausschließlich der Kern der Fuzzy-Ergebniszugehörigkeitsfunktion, d.h. die Menge der Elemente mit dem höchsten Erfüllungsgrad, betrachtet. Man kann zwischen First of Maximum (FOM), Middle of Maximum (MOM) und Last of Maximum (LOM) unterscheiden, wobei der kleinste Wert des Kerns, sein Median oder sein höchster Wert crisp zurückgegeben wird. (van Leekwijck & Kerre 1999, S. 166) Das wahrscheinlich bekannteste und weit verbreitetste Defuzzifizierungs-Verfahren ist die Center of Gravity (COG) Methode. Sie sieht die Ergebniszugehörigkeitsfunktion als Verteilung an und gibt den gewichteten Durchschnitt dieser, den Schwerpunkt ihrer Fläche entlang der x-Achse, zurück. (van Leekwijck & Kerre 1999, S. 167)

Die Resultate der Anwendung der oben genannten Defuzzifizierungsstrategien sind in Abbildung 2.4 skizziert.

## 3. Ziele und Anforderungen eines XAI-Empfehlungssystems

In diesem Kapitel wird die Forschungsfrage aus der aktuellen Problematik bei der Auswahl und Anwendung von XAI-Methoden abgeleitet. Zudem werden Ziele und Anforderungen definiert, die es für die Erstellung eines Lösungsansatzes zu erfüllen gilt.

### 3.1. Ableitung der Forschungsfragestellung

Wie in Kapitel 2.2.2 identifiziert, gibt es einige Faktoren, die die Auswahl und anschließende Anwendung von geeigneten XAI-Methoden für einen Nutzer als schwierig gestalten: Der Aufwand bei der Informationsbeschaffung innerhalb des segmentierten und sich schnell entwickelnden Forschungsbereichs der XAI ist erheblich. Außerdem mangelt es an Unterstützung bei der Operationalisierung von XAI-Methoden, die sich für den jeweiligen Anwendungskontext eignen.

Ein Empfehlungssystem kann diese Probleme beheben. Das darin gebündelte XAI-Expertenwissen kann jederzeit und schnell abgerufen werden und eine benutzerspezifische, begründete Empfehlung einer Methode erstellt werden. Um der Forschungsfrage nachzugehen und den Nutzer bei der Auswahl und anschließenden Anwendung von geeigneten XAI-Methoden auf Black-Box Modelle durch Empfehlungen zu unterstützen, wird daher ein XAI-Empfehlungssystem implementiert, das nachfolgend XAIR (XAI Recommender) genannt wird.

Für die Erstellung des Systems gilt es die folgenden Teilfragen zu klären:

- Wie wird die Eignung von XAI-Methoden beurteilt?
- Auf welche Kriterien eines spezifischen Modell-, Daten- und Nutzungskontexts muss bei der Auswahl und Eignungsbeurteilung von XAI-Methoden geachtet werden?
- Wie lässt sich das Wissen der Beurteilung der XAI-Methodeneignung hinsichtlich der identifizierten Kriterien formalisieren?

Vor der Beantwortung der oben genannten Fragen werden zunächst die Ziele und die für das Erreichen der Ziele notwendigen Anforderungen an den XAIR ermittelt.

### 3.2. Ziele und Annahmen des XAI-Empfehlungssystems

Das Ziel des XAIR ist das Erleichtern der Auswahl und Anwendung von XAI-Methoden für einen Nutzer ohne tiefe ML-Kenntnisse durch den Erhalt begründeter XAI-Methodenempfehlungen. Durch Vorschläge konkreter Implementierungen und technischer sowie methodischer Aspekte soll der Nutzer

nicht nur bei der Auswahl der XAI-Methode unterstützt, sondern auch zu ihrer anschließenden Anwendung motiviert werden. Eine zusammenfassende Aufführung der Richtlinien und Prinzipien zur Förderung von Transparenz (siehe Kapitel 2.2.2) soll dem Nutzer als Orientierung zur nachfolgenden bzw. allgemeinen Verwendung von XAI dienen.

Für eine Methodenempfehlung werden eignungsbeeinflussende Eigenschaften des Modell-, Daten- und Nutzungskontexts berücksichtigt, die in einer solchen Form in keiner existierenden Literatur erfasst sind. Zudem trägt es als erstes Empfehlungs- und Expertensystem im Kontext von XAI, durch Organisation und Formalisierung des verstreuten Fachwissens, zum aktuellen Stand der Forschung bei.

Die Empfehlungen können für ein bereits implementiertes Modell eingeholt werden, welches ein Black-Box Modell sein kann. Es wird angenommen, dass dieses Klassifikations- oder Regressionsaufgaben mithilfe von tabellarischen Daten löst.

Die Zielgruppe des XAIRs besteht aus den Modellentwicklern, die datenwissenschaftliche und -analytische Kenntnisse besitzen, und den Personen, die für das ML-Modell und seinen Freigabeprozess verantwortlich sind. Dabei wird davon ausgegangen, dass diese über Domänenwissen verfügen und ML-Fachtermini verstehen, aber nicht zwangsläufig tiefere ML-Kenntnisse vorweisen. Für eine Anwendung wird außerdem Wissen über den Nutzungskontext und Eigenschaften des Trainingsdatensatzes des zu erklärenden ML-Modells vorausgesetzt.

### **3.3. Anforderungsanalyse**

Im Rahmen von Requirement Driven DSR bieten Anforderungen methodische Unterstützung bei der Entwicklung eines Artefakts: Vor der Entwurfsphase dienen sie der Problem- und Domänenbeschreibung und als Grundlage für Designentscheidungen, während und nach dem Entwurf als Fortschrittsdokumentation und Entscheidungsbegründung. Eine Betrachtung von Anforderungen als Repräsentationen entworfenen Konstrukte ermöglicht außerdem frühe (iterative und inkrementelle) Evaluationen auf Basis von diesen. (Braun et al. 2015)

Daher wird initial eine Anforderungsanalyse für den XAIR durchgeführt.

#### **3.3.1. Aufbau und Durchführung**

Für den XAIR werden typische nicht-funktionale Software-Anforderungen, angelehnt an die Softwarequalitätskriterien nach Boehm et al. (1976), aufgenommen und mittels Brainstorming funktionale Anforderungen an ein Empfehlungssystem ermittelt.

Außerdem werden diese um Anforderungen ergänzt, die in halbstrukturierten Online-Einzelinterviews erhoben werden. Diese offene, qualitative Interviewmethode eignet sich für problemzentrierte Befragungen, zur Prüfung bestehender Hypothesen und zur Gewinnung weiterer Einsichten (Buber & Holz-müller 2009, S. 465). Ein halbstrukturierter Interviewleitfaden, welcher in Anhang A.1 zu finden ist, gewährleistet dabei Kohärenz und Vergleichbarkeit der Aussagen, bietet allerdings auch Raum für natürliche Konversation und eventuelle Ergänzungsfragen.

Diese Anforderungserhebung in Form von Interviews findet mit fünf Personen statt, welche die Zielgruppe der Entscheidungshilfe widerspiegeln. Darunter befinden sich:

- eine Person aus dem Data Science Kontext, die mit XAI vertraut ist (Person A)
- mehrere Personen mit wenigen XAI-Kenntnissen, die
  - ausschließlich tiefe IT-Kenntnisse aufweist (Person B) und aufgrund diverser Erfahrungen die Rolle eines Modellverantwortlichen (z.B. eines SCRUM Product Owners) einnimmt
  - Data Science und ML-Erfahrung in der Forschung und nicht in der Industrie aufweisen und wenig (Person C) bzw. gar nicht (Person D) mit XAI vertraut sind
- eine Person mit wenigen IT- und nur theoretischen, nicht innerhalb eines Unternehmens angewandten ML-Kenntnissen (Person E)

### 3.3.2. Erhobener Anforderungskatalog

Die gesammelten und zusammengefassten Anforderungen werden nachfolgend, aufgeteilt in funktional (FR) und nicht-funktional (NFR), vorgestellt.

#### Funktionale Anforderungen

##### FR1 Übersichtliche Darstellung des Empfehlungsergebnisses

Das System soll die empfohlenen XAI-Methoden übersichtlich, kompakt mit „zwei, drei Kernaussagen“ (Person B) und nach Eignung geordnet darstellen. Eine Ordnung und die Angabe des Unterschieds zwischen den Methodeneignungen beim Empfehlungsergebnis werden von den Befragten als positiv wahrgenommen. Deshalb soll die Differenz der Gütebewertungen der vorgeschlagenen Methoden ersichtlich und ihre Eignung somit vergleichbar sein.

##### FR2 Nachvollziehbare Begründung der Empfehlungsentscheidung

Das System soll das Empfehlungsergebnis einfach, nachvollziehbar und transparent begründen. Die Begründung soll dabei, neben den eigentlichen Empfehlungsergebnissen, schnell ersichtlich sein, da für Person C bei routinierter Anwendung nur noch dieser Teil interessant wäre. Bei einer gelegentlichen Nutzung wäre er aufgrund des Lerneffekts ebenfalls wichtig. Im Gegensatz dazu möchte Person B die Entscheidungsbegründung am liebsten ausblenden können, wobei sie sich im Sinne der Akzeptanz und der Wiederverwendbarkeit des Systems der Wichtigkeit seiner Nachvollziehbarkeit bewusst ist. Daher wird eine optische Trennung des Empfehlungsergebnisses von seiner Begründung angestrebt. Allen Befragten ist es wichtig zu wissen, inwiefern ihre Eingaben für die Methodenauswahl relevant waren. Daher soll die Eignung der einzelnen Methoden auf Grundlage dieser kurz aufgezeigt werden, was einen mehrfach gewünschten „direkten Vergleich mit verschiedenen Vor- und Nachteilen [...] z.B. hinsichtlich verschiedener Kriterien“ (Person E), ermöglicht. Person D fände es gut, wenn „auch die Frage beantwortet wird, warum die anderen Methoden nicht empfohlen werden“. Um eine fundierte Entscheidung zu treffen, würde sie dafür auch in Kauf nehmen, dass eine vergleichende Übersicht aller XAI-Methoden etwas komplexer ist.

Zur Förderung der Nachvollziehbarkeit soll außerdem eine Erklärung der Systemfunktionsweise verfügbar sein.

**FR3 Eingabe benötigter Parameter**

Die Eingabe und frühzeitige Validierung der benötigten Parameter soll über eine grafische Benutzeroberfläche (GUI) erfolgen.

Person D würde die Eingabeparameter gerne weitestgehend automatisiert aus den Trainingsdaten ermitteln lassen. Allerdings dürfen sensible Daten, abhängig von ihrem Verwendungszweck und vom Betriebskontext des XAIRs, aufgrund geltender Datenschutzrichtlinien nicht über eine öffentliche URL weitergegeben werden. Daher soll die Eingabe manuell erfolgen und ohne die Bereitstellung der Trainingsdaten möglich sein. Die Befragten waren sich einig, dass die Eingaben auf ein Minimum beschränkt sein sollten, um einen erhöhten Aufwand bei der Anwendung zu vermeiden. Ein höherer initialer Konfigurationsaufwand wird von Person A akzeptiert, wenn das Empfehlungssystem, ohne weiteren Aufwand, wiederholbar in einer ML-Pipeline eingesetzt werden kann.

**FR4 Berücksichtigung nutzerspezifischer Präferenzen**

Zur Steigerung der Nutzerzufriedenheit soll das System Präferenzen des Nutzers für die Empfehlung berücksichtigen. Im Hinblick auf eine beschränkte Verfügbarkeit von Ressourcen wurde die Leistung der XAI-Methoden von allen Interviewpartnern als gewünschtes Eingabemerkmal genannt. Person A wünscht sich außerdem die Angabe eines spezifischen Formates der resultierenden Erklärung. Dies kann damit begründet werden, dass sie über tiefere Kenntnisse von XAI-Methoden, ihren Ergebnisformaten und -interpretationen verfügt. Den anderen Teilnehmern ist, mit den Worten von Person B, das „Ich-Kanns-Verstehen wichtiger als das Format“. Diese Präferenz wurde nachfolgend als „Angabe der Zielgruppe/Komplexität der Erklärung“ formuliert, womit die Interpretationsschwierigkeit des Methodenergebnisses angepasst und eine Nachvollziehbarkeit gewährleistet werden kann.

Die ermittelten, nutzerspezifischen Präferenzen bzgl. der vorgeschlagenen XAI-Methoden, die das System entgegennehmen soll, sind nachfolgend zusammengefasst aufgelistet:

- Leistung/Laufzeit der XAI-Methode
- Präferiertes Format der Erklärung
  - Umfang der Erklärung (Globale oder lokale Erklärungen)
  - Resultierendes Erklärungsformat (bspw. textuell, grafisch, regelbasiert)
- Angabe der Zielgruppe/Komplexität der Erklärung
- Personeller Aufwand bei der Einarbeitung und Implementierung der XAI-Methode

**FR5 Detaillierte Informationen zu empfohlenen Methoden**

Der Nutzer soll die Möglichkeit haben, detailliertere Informationen über die empfohlene(n) Methode(n) zu erhalten. Um der Empfehlung trauen zu können würde Person D die Zeit investieren und sich die empfohlene XAI-Methode und ihre Quellen mit tieferem Wissen im Detail ansehen. Auch Person A findet nähere Informationen zur Funktionsweise und den Ergebnissen der Methode, sowie Vorschläge für eine anschließende Implementierung, sehr gut.

Für Person B ist allerdings „die Empfehlung das Wichtigste und diese weiteren Informationen wären [...] sekundär, gegebenenfalls hilfreich“. Daher sollen die spezifischen Methodeninformationen bei Desinteresse übersprungen werden können.

**FR6 Speicherbarkeit**

Das System sollte dem Nutzer die Möglichkeit geben, die Eingaben und das Empfehlungsergebnis abzuspeichern und wieder aufzurufen.

**FR7 Nachträgliche Änderbarkeit der Eingabedaten**

Nach Erhalt des Ergebnisses soll dem Nutzer die Möglichkeit der Änderung der Eingabedaten und des erneuten Ausführens des XAIRs geboten werden.

**FR8 Förderung der Bereitschaft der Anwendung von XAI**

Für die vorgeschlagene XAI-Methode soll eine Empfehlung bzgl. einer Implementierung gegeben werden, um den Rechercheaufwand der Auswahl zu reduzieren. Zusätzliche Hinweise und Vorschläge für mögliche Konfigurationsparameter sollen den Einarbeitungsaufwand reduzieren. Das System soll außerdem Empfehlungen zum weiteren Vorgehen geben.

**Nicht-Funktionale Anforderungen****NFR1 Portabilität**

Das System soll dediziert verwendbar sein, soll allerdings minimal-invasiv und ohne große codeseitige Anpassungen in bestehende ML-Pipelines integriert werden können.

**NFR2 Benutzbarkeit**

Die manuelle Eingabe der benötigten Daten soll, ebenso wie die Präsentation des Empfehlungsergebnisses, über eine grafische Weboberfläche erfolgen.

**NFR3 Benutzerfreundlichkeit**

Das System soll eine leicht verständliche und intuitive Benutzeroberfläche bieten, sodass eine Bedienung ohne größeren Einarbeitungsaufwand möglich ist.

**NFR4 Wartbarkeit und Skalierbarkeit**

Das System soll durch einen modularen Aufbau leicht wartbar sein und erweiterbar hinsichtlich weiterer

- XAI-Methoden
- (Nutzerspezifischer) Eingabekriterien
- Expertenmeinungen bzgl. der Methodenbewertungen

**NFR5 Effizienz**

Eine sehr hohe Berechnungsdauer zur Ermittlung des Empfehlungsergebnisses soll vermieden werden. Für Person B wäre ein verzögerter, asynchroner Ergebniserhalt jedoch „in Ordnung, wenn [Person B] nachvollziehen kann, dass da ein gewisser Rechenaufwand [...] dahintersteckt“.

**NFR6 Konsistenz/ Nicht-Volatilität der Ergebnisse**

Die Ergebnisse des Systems sollen konsistent, reproduzierbar und somit zuverlässig sein. Minimale Än-

derungen der Eingaben sollten keine großen Änderungen in der Empfehlungsentscheidung verursachen.

Außer NFR5 werden von den Befragten keine nicht-funktionalen Anforderungen genannt. Lediglich Person A, die sich mit der Bereitstellung von Infrastrukturen für ML-Pipelines beschäftigt, und Person D legen neben der Effizienz auch Wert auf eine leichte Integration und Einsetzbarkeit des Systems in einer ML-Pipeline (NFR1). NFR1 wird umgesetzt und eine Integrierbarkeit in die verbreitete End-to-End Plattform Kubeflow Pipelines angestrebt. Die Plattform erleichtert das Erstellen und den Einsatz von portablen, skalierbaren und auf Containern basierenden Python ML-Pipelines (Kubeflow 2020). Mit NFR1 soll der dadurch automatisiert ausgeführte Schritt der Auswahl einer XAI-Methode obligatorisch gemacht, und der Modellverantwortliche somit zu einer Auseinandersetzung mit XAI bewegt werden.

Die Nutzer eines Empfehlungssystems wünschen sich häufig eine interaktive Einflussnahme auf die Empfehlungsgenerierung und -darstellung, was bspw. durch das Kritisieren einzelner Empfehlungen oder die Auswahl oder Gewichtung von Datenquellen möglich ist (Ziegler & Loepp 2019, S. 20). Auch wenn eine Integration von Nutzer-Feedback bzgl. der XAI-Methodeneignung längerfristig angestrebt wird und der Wunsch nach einer Gewichtung der nutzerspezifischen Eingabepreferenzen auch von Person A geäußert wird, ist eine Umsetzung dieser Interaktivität für den im Rahmen dieser Arbeit entwickelten Prototypen nicht vorgesehen.

Nachfolgend wird ein Überblick über die für das Software-Artefakt aufgenommenen Anforderungen gegeben.

### 3.3.3. Anforderungsübersicht

#### Funktionale Anforderungen

FR1 Übersichtliche Darstellung des Empfehlungsergebnisses

FR1.1 Vergleichbarkeit der Eignung der XAI-Methoden

FR2 Nachvollziehbare Begründung der Empfehlungsentscheidung anhand der Eingaben

FR2.1 Eignungsvergleich der Methoden pro Eingabeparameter

FR2.2 Erklärung der Funktionsweise des Empfehlungssystems

FR3 Eingabe benötigter Parameter

FR3.1 Eingabe der benötigten Parameter über Weboberfläche

FR3.2 Validierung der Eingabeparameter

FR3.3 Möglichkeit der Ausführung ohne Preisgabe der Daten

FR3.4 Beschränkung der Eingaben auf ein Minimum

FR4 Berücksichtigung nutzerspezifischer Präferenzen

FR4.1 Leistung/Laufzeit

FR4.2 Präferiertes Erklärungsformat

FR4.3 Komplexität/Zielgruppe der Erklärung

FR4.4 Personeller Aufwand bei der Vorbereitung (Einarbeitung/Implementierung)

FR5 Detaillierte Informationen zu empfohlenen Methoden

FR5.1 Optional überspringbar

FR6 Speicherbarkeit der Empfehlungsergebnisse

FR7 Nachträgliche Änderbarkeit der Eingabedaten

FR8 Förderung der Bereitschaft der Anwendung von XAI

FR8.1 Empfehlung einer Implementierung

FR8.2 Empfehlungen zum weiteren Vorgehen

### **Nicht-Funktionale Anforderungen**

NFR1 Portabilität

NFR1.1 Minimal-invasive Integration in ML-Pipeline

NFR2 Benutzbarkeit

NFR3 Benutzerfreundlichkeit

NFR4 Wartbarkeit und Skalierbarkeit

NFR5 Effizienz

NFR6 Konsistenz/ Nicht-Volatilität der Ergebnisse



## 4. Konzeption des XAI-Empfehlungssystems

Dieses Kapitel geht auf den Entwurf des XAI-Empfehlungssystems ein, der die Grundlage der Implementierung des XAIRs schafft. Davor werden die im Kontext der Forschungsfrage relevanten Teilfragen (siehe Kapitel 3.1) geklärt.

### 4.1. Identifikation geeigneter XAI-Methoden

Um dem Nutzer geeignete XAI-Methoden vorschlagen zu können, muss zunächst geklärt werden, was unter der Eignung einer XAI-Methode verstanden wird bzw. wann diese als geeignet gilt.

#### 4.1.1. Beurteilung der Eignung einer XAI-Methode

Das Ziel der Anwendung einer XAI-Methode ist der Erhalt einer Erklärung, die dem Nutzer die Funktionsweise eines ML-Modells erklärt oder ihm Gründe für bestimmte Modell-Entscheidungen liefert (Kraus et al. 2021, S. 24).

Die Beurteilung einer XAI-Methode anhand ihres Erklärungsergebnisses gestaltet sich allerdings als schwierig, da Erklärungen nicht nur die Darstellung von Zusammenhängen und Ursachen (Kausalattribution) sind, sondern kontextbezogen und durch soziale Überzeugungen und kognitive Verzerrungen sehr subjektiv (Miller 2019). Die Bereiche der Philosophie, Psychologie und Sozialwissenschaften beschäftigen sich schon lange mit der Frage, was Erklärbarkeit ist, wie sie formalisierbar ist, welche Eigenschaften Erklärungen effektiv und verständlich für Nicht-Experten machen und wie man ihre Qualität messen kann (Doshi-Velez & Kim 2017). Das Problem der Quantifizierung des Verständlichkeitsgrads einer Erklärung für den Menschen wurde noch nicht gelöst, daher ist eine (mathematische) Formalisierung und Messung nicht möglich (Guidotti et al. 2018).

Im Rahmen von XAI soll laut Leslie (2019) eine resultierende Erklärung die Interpretationsbedürfnisse des Nutzers erfüllen, die den Kontext der Modellanwendung, die potenziellen Auswirkungen und die domänenspezifischen Anforderungen berücksichtigen. Abhängig von der Domäne müssen außerdem mögliche Erklärungsstandards oder Begründungs-Vergleichsmaßstäbe eingehalten werden.

Da eine generalisierte Anwendbarkeit des XAIRs angestrebt wird, kann die Güte einer Erklärung nicht anhand des schwer messbaren sozialen und kognitiven Kontexts oder aufgrund domänenspezifischer Standards ausgemacht werden. Eine von Doshi-Velez & Kim (2017) vorgeschlagene Taxonomie für die Evaluierung der Interpretierbarkeit auf Anwendungs-, Nutzer- und Funktionsebene kommt daher nicht infrage, da sie Erklärung im Kontext spezifischer ML-Aufgaben evaluieren.

Neben der Qualität ihres Ergebnisses kann die Eignung bzw. Güte einer XAI-Methode anhand ihrer Voraussetzungen und der Einfachheit ihrer Anwendung im jeweiligen Modell-, Daten- und Nutzungskontext beurteilt werden.

Wenn eine XAI-Methode durch nicht erfüllte Voraussetzungen nicht verwendbar ist, ist sie ungeeignet. Das ist u.a. abhängig von dem beabsichtigten algorithmischen Ansatz (Leslie 2019) und folglich auch von dem Modell, das durch die Anwendung des Algorithmus auf die Trainingsdaten entsteht.

Ist eine grundsätzliche Ausführbarkeit gegeben, ist eine XAI-Methode aufgrund ihrer Funktionsweise abhängig von den Datenformaten oder -eigenschaften mehr oder weniger gut geeignet.

Eine limitierte Verfügbarkeit von Hardware-Ressourcen reduziert zudem die Eignung von XAI-Methoden, da diese teilweise sehr hohe Leistungsanforderungen haben. Allerdings ist die Höhe von personellen und Hardware-Ressourcen kontextabhängig, da kritische Kontexte einen höheren Ressourcenaufwand für die Anwendung der Interpretierbarkeitmethoden rechtfertigen (Leslie 2019).

Außerdem sind die Bedürfnisse des Nutzers, der nicht nur die Erklärung konsumiert, sondern die Methode auch anwendet, für die Beurteilung der Güte einer XAI-Methode relevant. Eine sehr gute Methode zeichnet sich durch leichte Anwendbarkeit ohne großen Einarbeitungsaufwand aus, was allerdings erneut aufgrund der Subjektivität nicht metrisch messbar ist.

Zusammenfassend lässt sich die Güte einer XAI-Methode nicht aufgrund der Qualität der resultierenden Erklärung beurteilen, sondern anhand von Eigenschaften des Modell-, Daten- und Nutzungskontexts. Sie wird demnach anhand von Faktoren ermittelt, die

- die Anwendung der XAI-Methode erschweren oder unmöglich machen,
- aufgrund der algorithmischen Beschaffenheit der XAI-Methode einen negativen, verfälschenden Einfluss auf ein solides, kohärentes und vernünftiges Erklärungsergebnis haben,
- die Interpretierbarkeit der Erklärung mindern oder verkomplizieren.

Nach einer initialen Auswahl von XAI-Methoden werden anhand dieser konkrete Kriterien mit den oben genannten Eigenschaften identifiziert, welche für die einzelnen XAI-Methoden mehr oder weniger relevant sind.

#### 4.1.2. Eingrenzung der XAI-Methoden und Auswahl der Implementierungen

Es besteht ein allgemeiner Konsens darüber, dass eine Kombination mehrerer Erklärungstechniken förderlich für den Erhalt eines besseren Gesamtbilds des Modells und einer vollständigeren Erklärung ist (vgl. Belle & Papantonis (2020), Leslie (2019), Kangur (2020), Gilpin et al. (2019)). Sofern die Anwendung eines transparenten Modells, welches bevorzugt werden sollte (Leslie 2019, S. 45), aufgrund einer schlechten Leistung für die Erfüllung der Aufgabe nicht infrage kommt, empfehlen Kangur (2020) und Belle & Papantonis (2020) in ihren Richtlinien die Anwendung von:

1. *Visualisierungstechniken* zur Vermeidung von Verzerrungen bereits bei der Datenexploration (Kangur 2020). Eine Verwendung als ergänzende Technik zur Verdeutlichung der Entscheidungsgrenzen wichtiger Features wird von Belle & Papantonis (2020) empfohlen.
2. *Lokalen Methoden*, die am besten für richtig und falsch vorhergesagte Dateninstanzen angewendet werden (Kangur 2020), um zu sehen, wie sich kleine Veränderungen auf das Ergebnis auswirken.

Das kann ggf. Aufschlüsse über mögliche Verbesserungen des Feature Engineerings geben (Kangur 2020). Bei der Abdeckung der angewendeten XAI-Methoden und des Erklärungsumfangs (ob global oder lokal) sollte laut Leslie (2019) „local-first“ gedacht werden.

3. *Globalen Erklärungen*, entweder durch Modellvereinfachung oder Feature Relevanz Methoden, um Aufschluss über die Wichtigkeit bestimmter Features für das Gesamtsystem zu erhalten. Diese Ergebnisse sollten laut (Kangur 2020) wenn möglich zusammen mit den Metriken der Modell-Leistung in Kenntnis genommen werden.

Auch die Umfragen zum Einsatz von XAI in Unternehmen verschiedener Bereiche von Bhatt et al. (2020) zeigen, dass die am häufigsten genutzten Verfahren Feature Relevanz Methoden sind, und auch *Counterfactual Explanations* („Kontrafaktische Erklärungen“) im Deployment eingesetzt werden.

Angelehnt an diese Richtlinien und Umfrageergebnisse werden konkrete Methoden dieser Erklärungsverfahren anhand ihrer Aktualität und der Häufigkeit ihrer Erwähnungen in zusammenfassenden wissenschaftlichen Veröffentlichungen ausgewählt (u.a. Molnar (2019), Carvalho et al. (2019), Belle & Papantonis (2020), Hall & Gill (2019), Adadi & Berrada (2018)).

Dabei werden nur post-hoc anwendbare XAI-Verfahren in die initiale Auswahl aufgenommen, da angenommen wird, dass das zu erklärende Modell bereits entwickelt ist. Auch wenn dieses nicht zwangsläufig eine Black-Box ist, werden modellspezifische Methoden für den XAIR Prototyp bewusst nicht berücksichtigt. So werden spärliche oder leere Empfehlungsergebnismengen anwendbarer XAI-Methoden vermieden.

Um die Bereitschaft der tatsächlichen Anwendung der XAI-Methoden zu erhöhen (Anforderung FR5), wird für jede Methode eine Implementierung ausgewählt, die dem Nutzer zusätzlich mit weiteren Hinweisen zur Anwendung vorgeschlagen wird. Diese Implementierungen werden im Januar 2021 evaluiert und final für das Empfehlungssystem ausgewählt. Andere, die die vorhandenen Probleme der Ausgewählten beheben und sich daher besser für eine Empfehlung eignen, werden nicht mehr in Betracht gezogen.

Für die Auswahl werden ausschließlich Implementierungen der Programmiersprache Python (Python Software Foundation 2020) berücksichtigt. Das wird damit begründet, dass die Sprache im Data Science Umfeld sehr weit verbreitet ist und dass die Implementierungen somit, im Hinblick auf einen automatisierten Einsatz, einfach als funktionsbasierte Komponenten in eine Kubeflow Pipeline integrierbar sind. Dabei wird der Code einer Komponente aus einer Python-Funktion generiert und die notwendige Komponentenspezifikation automatisch erstellt (vgl. Kubeflow (2021)).

Die ausgewählten Implementierungen unterliegen keiner Copyleft-Lizenz, sodass eine eventuell resultierende Kubeflow XAI-Komponente nicht unter den gleichen, freien Lizenzbedingungen veröffentlicht werden muss. Weitere Auswahlkriterien sind unter anderem eine öffentliche Bereitstellung, bspw. auf GitHub, und eine hohe Aktivität/Wartung der Projektarchive, da man von einem häufig verwendeten und von vielen Nutzern gewarteten Projekt eine gewisse Langlebigkeit und Fehlerreduktion erwarten kann.

Die Bewertungen gesichteter Implementierungen hinsichtlich aller Kriterien sind in Anhang A.2 zu finden.

Generelle Probleme bei der Suche und Auswahl der Implementierungen sind, dass einige XAI-Methoden teilweise ausschließlich in R und nicht in Python umgesetzt sind, dass viele unzureichend dokumentiert und kommentiert sind und dass verschiedene Implementierungen desselben Verfahrens verschiedene Schwächen aufweisen. Es gibt außerdem nur wenige XAI Ansätze, die mit effizienten Implementierungen einhergehen, was damit begründet werden kann, dass dieses Gebiet noch jung und im Entstehen ist (Belle & Papantonis 2020, S. 20).

Nachfolgend werden die ausgewählten XAI-Methoden und ihre Implementierungen vorgestellt.

## Visualisierungstechniken

### Partial Dependence Plot (PDP) und Individual Conditional Expectation (ICE)

PDP und ICE erklären beide visuell die Entscheidung des Modells anhand der marginalen Einflüsse einzelner Eingabe-Features und beantwortet somit die Frage: Wie ist der Zusammenhang zwischen dem betrachteten, (unabhängigen) Feature und der Vorhersage? (Molnar 2019) Der Unterschied zwischen diesen Visualisierungsmethoden ist, dass die resultierende Erklärung bei PDP global ist, bspw. „Welchen durchschnittlichen Einfluss hat das Alter auf die Bewilligung eines Kredits durch die Bank?“, während sich die von ICE auf eine Dateninstanz bezieht („Welchen Einfluss hat *mein* Alter auf die Bewilligung *meines* Kredits durch die Bank?“). PDP ist außerdem für die Darstellung kombinierter Effekte zweier Features, d.h. ihres Gesamteffekts und ihren Interaktionseffekten mit anderen Features, geeignet (Molnar 2019).

Die Methoden funktionieren nach demselben Prinzip: Um den marginalen Effekt eines zu betrachtenden Features zu ermitteln, werden Vorhersagen für Dateninstanzen gemacht, bei denen dieser Feature-Wert verändert wird, während die aller anderen unverändert bleiben. Das gibt Aufschluss darüber, wie sich die Vorhersage ändert, wenn sich das betrachtete Feature ändert. (Molnar 2019)

ICE stellt die Veränderung der Vorhersage durch das Feature für jede Dateninstanz als eine Linie dar, wohingegen das Ergebnis eines PDP der Durchschnitt aller ICE Plots ist (Goldstein et al. 2015).

Bei der Verwendung der Methoden ist allerdings eine Unabhängigkeit der Eingabe-Features vorausgesetzt. Sollten sie untereinander korrelieren, kann es zur Generierung seltener, unrealistischer oder sogar unmöglicher Dateninstanzen kommen. Ein Beispiel hierfür anhand der Dateninstanz einer Person ist die Veränderung des Features „Gewicht“ auf „48 kg“, wenn das Feature „Körpergröße“ den Wert „200 cm“ behält. (Molnar 2019)

Solche neu erzeugten Dateninstanzen sind extrapoliert, d.h. sie liegen außerhalb der Trainingsdatenverteilung, und sind durch ihre unüblichen Feature-Kombinationen sehr unwahrscheinlich. Beide Verfahren berücksichtigen diese jedoch bei der Berechnung des Ergebnisses, wodurch es verfälschend beeinflusst werden kann (Goldstein et al. 2015).

PDP und ICE werden in der Praxis häufig zusammen angewendet, da ICE durch starke Korrelationen der Features verursachte Ungenauigkeiten von PDP aufdecken kann und da bei einer ausschließlichen Verwendung von PDP heterogene Effekte der Daten unerkannt bleiben (Goldstein et al. 2015). Durch Visualisierung des Durchschnitts heben sich gleich viele positive und negative Auswirkungen eines

Features gegenseitig auf, weshalb ICE Plots für einzelne Instanzen genauer betrachtet werden sollten (Molnar 2019).

*Implementierung, vgl. Tabelle A.1*

Die vorhandenen Python Implementierungen unterstützen jeweils beide Visualisierungsverfahren und geben PDP und ICE oft in demselben Plot aus. Für den XAIR wird die Implementierung *PDPbox* ausgewählt, da sie im Gegensatz zu einer anderen in Betracht gezogenen Alternative die Visualisierung nicht numerischer Features erlaubt. *PDPbox* überzeugt durch eine gute Community, eine ansprechende grafische Visualisierung und das Bereitstellen verschiedener Variationen von PDP Interaktions-Plots. Zudem bietet es diverse zusätzliche Informationsgrafiken zur Analyse der Auswirkung eines Features auf die Vorhersage. Unter anderem kann die durchschnittliche Vorhersage für einen bzw. zwei Feature-Werte und die Verteilung der tatsächlichen Vorhersagen für diese Feature-Werte dargestellt werden. Die Werte numerischer Features sind dabei Wertebereiche (Perzentile).

### Accumulated Local Effects (ALE)

Die Accumulated Local Effects (ALE) liefern ebenfalls globale, auf den Wichtigkeiten der Features basierende Visualisierungen. Sie beantworten dieselbe Frage wie PDPs, mit dem Unterschied, dass auch Features betrachtet werden können, die eventuell mit anderen korrelieren.

Für die Berechnung des Einflusses des Features wird dessen Wertebereich zuerst in lokale Bereiche unterteilt. Für jeden Bereich wird der Feature-Wert zwischen den Rändern bewegt und der durchschnittliche Unterschied der Vorhersage berechnet: Was sagt das Modell vorher, wenn man das Feature innerhalb eines kleinen Intervalls um seinen Wert ändert? (Molnar 2019)

Eine Lokalisierung der Feature-Werte mithilfe dieser Bereiche vermeidet eine Extrapolation und somit die Erstellung unrealistischer Dateninstanzen aufgrund von Korrelation der Eingabe-Features (Goldstein et al. 2015). Die Summe dieser arithmetischen Mittel wird zentriert dargestellt ( $\Sigma \bar{x} = 0$ ). Die Visualisierung zeigt somit die (globale) Auswirkung des Feature-Werts auf die Vorhersage im Vergleich zu der durchschnittlichen Vorhersage. (Molnar 2019)

Wenn also die Vorhersage  $\hat{y} = -500$  (Kredithöhe in Euro) bei Eingabe von  $x = 26$  (Alter in Jahren), dann ist die vorhergesagte Kredithöhe für eine 26-jährige Person 500 Euro niedriger als die durchschnittliche Kredithöhe.

Die Berechnung des Einflusses zweier Features ist auch möglich. Der Graph zeigt aber im Gegensatz zu PDP ausschließlich den Effekt zweiter Ordnung an, d.h. die zusätzliche Auswirkung der Interaktion der beiden Features auf die Vorhersage. (Molnar 2019)

*Implementierung, vgl. Tabelle A.2*

Die Auswahl guter Python-Implementierungen für ALE ist klein, die Unterstützung durch eine große Community selten gegeben und die Dokumentation oft unzureichend. Trotz einer kleinen Community auf GitHub fällt die Entscheidung auf *PyALE*, da sie als einzige Implementierung die Visualisierung kategorischer Features unterstützt.

## Lokale Methoden

### SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) ist ein spieltheoretischer Ansatz, der Shapley Werte mit lokalen, linearen Regressionsmodellen verbindet (Lundberg & Lee 2017).

Laut der Studie von Bhatt et al. (2020) werden Feature Relevanz Verfahren am häufigsten von Unternehmen genutzt. Darunter fällt die am meisten verwendete Methode der Shapley Werte. Sie beantwortet die Frage, welchen Beitrag der Wert eines oder mehrerer Features zur Vorhersage liefert, verglichen mit der durchschnittlichen Vorhersage. Das basiert auf der Idee, dass das Vorhersageergebnis fair unter allen Features aufgeteilt wird und dass für die Bestimmung der Wichtigkeit eines einzelnen Features, alle möglichen Feature-Kombinationen berücksichtigt werden sollten. (Molnar 2019)

SHAP gibt lokal Aufschluss über die Feature Relevanzen, z.B.: „Inwieweit wurde meine Vorhersage für die Kredithöhe durch die Tatsache beeinflusst, dass ich vier Bankkonten habe, statt nur der durchschnittlichen Anzahl von zwei Konten?“

Für ihre Berechnung werden zunächst die Vorhersagen aller möglichen Feature-Kombinationen einer Dateninstanz berechnet. Fehlt in einer solchen Kombination ein Feature, wird sein Wert durch den einer anderen Dateninstanz ersetzt (permutiert). Anschließend werden die Differenzen der Werte mit und ohne den zu betrachtenden Feature-Wert gewichtet aggregiert, um den marginalen Beitrag des Features für das Ergebnis zu erhalten. (Lundberg & Lee 2017)

Kleine Kombinationen mit wenig nicht permutierten Features und große mit vielen, erhalten dabei höhere Gewichte, da sie mehr zu den Haupt- und Gesamteffekten der Features aussagen als mittel-große. Die Ermittlung dieser additiven Wichtigkeiten der Features geschieht mit einem linearen Modell, dessen Koeffizienten den Shapley Werten entsprechen. (Molnar 2019)

*Implementierung, vgl. Tabelle A.3*

Lundberg & Lee (2017) stellen mit der Veröffentlichung ihres Papers auch eine Python Implementierung zur Verfügung, die neben der modellagnostischen KernelSHAP Methode auch modellspezifische Varianten, bspw. für Bäume oder DNNs, bietet. Allerdings wird die KernelSHAP Implementierung von *Alibi* dieser vorgezogen. Sie bietet einen Wrapper um die SHAP Implementierung von Lundberg & Lee (2017) und ermöglicht somit ein explizites Gruppieren kategorischer Variablen. Dadurch wird sichergestellt, dass ein Feature trotz Kodierungen weiterhin als eine Dimension betrachtet wird. Dies reduziert zudem die Berechnungszeit, da für ein Feature mit  $m$  Kategorien nur ein SHAP Wert statt  $m$  Werte berechnet werden muss (Coca 2020).

Beide getesteten Implementierungen verwenden die Visualisierungsfunktion von Lundberg & Lee (2017) und überzeugen durch grafisch ansprechende Darstellungen. Durch eine Aggregation einzelner Instanzvorhersagen ist auch der Erhalt globaler Visualisierungen möglich.

### Anchors

Obwohl die XAI-Methode Anchors nicht in der Umfrage von Bhatt et al. (2020) erwähnt wird, wird sie dennoch in die initiale Auswahl der XAI-Methoden des Empfehlungssystems aufgenommen. Sie ist ein populäres Verfahren und eine Weiterentwicklung der in der Literatur häufig erwähnten und

ebenfalls von Riberio et al. entwickelten Methode LIME (Ribeiro et al. 2016). Da durch die Aufnahme von Anchors für den Anfang genügend lokale XAI-Methoden in der Auswahl vorhanden sind, wird LIME zunächst nicht inkludiert und daher nicht weiter erläutert.

Anchors erklärt die Vorhersage eines Klassifikationsmodells lokal durch Extraktion von „Wenn-Dann“-Entscheidungsregeln, sogenannten Anchors, unter der Angabe der prozentualen Anwendbarkeit dieser auf andere Dateninstanzen und der Vorhersagegenauigkeit. (Riberio et al. 2018) Es beantwortet daher die Fragen: Nach welchen Regeln lässt sich die Vorhersage einer Instanz erklären? Auf wie viele Instanzen trifft die Regel zu („Coverage“) und wie gut beschreibt sie die Vorhersage der Instanz („Precision“)? (Molnar 2019)

Für eine zu erklärende Dateninstanz  $x$  werden durch Perturbationen neue Instanzen in ihrer Nachbarschaft generiert, die mit einer festgelegten Wahrscheinlichkeit (z.B. 95%) die gleiche Vorhersage erhalten wie  $x$ . Die Suche nach einem geeigneten Anchor Kandidaten findet mittels der Bottom-Up Strategie statt. Dabei wird eine leere Regel, die auf jede Dateninstanz zutrifft, iterativ um weitere Feature-Prädikate erweitert. Für jeden Kandidaten muss die Vorhersage durch das Modell berechnet werden. Da die Anzahl der potenziellen Anchors exponentiell zum Eingaberaum der Features ist, werden zur Reduktion der Berechnungskomplexität die geeignetsten Kandidaten von einem Multi-Armed-Bandit Algorithmus (KL-LUCB) ausgewählt und nur ihre Vorhersagen berechnet. (Riberio et al. 2018)

Mit Multi-Armed-Bandit Algorithmen können verschiedene Strategien durch sequentielle Auswahl effizient erkundet und ausgenutzt werden. Im Fall von Anchors wird ein Regelkandidat hinsichtlich der festgelegten Wahrscheinlichkeit evaluiert. (Molnar 2019)

Die vielversprechendsten Regeln werden in der nächsten Runde um weitere Feature-Prädikate ergänzt, bis die gewünschte Wahrscheinlichkeit erreicht wird. Sollten mehrere Anchors zur Erklärung der Instanz infrage kommen, wird der mit der höchsten Abdeckung gewählt. (Riberio et al. 2018)

*Implementierung, vgl. Tabelle A.4*

Für den XAIR wird die Implementierung von *Alibi* ausgewählt. Diese zeichnet sich im Gegensatz zur Implementierung der Veröffentlichung von Riberio et al. (2018) durch eine bessere Dokumentation, aussagekräftigere Code-Kommentare und durch eine höhere Konfigurierbarkeit aus.

### Counterfactual Explanations (CF)

Counterfactual Explanations (CF), kontrafaktische Erklärungen, liefern permutationsbasiert lokale Erklärungen für Klassifikationsaufgaben. Sie beantworten die Frage, welcher Feature-Wert (minimal) geändert werden muss, sodass sich die Vorhersage zu einer Vorgegebenen ändert (Molnar 2019), z.B. „Was muss ich ändern, damit die Bank meinen Kredit bewilligt?“. Durch ihre kontrastive Art sind sie leicht verständlich und von Menschen bevorzugt, da diese bspw. nicht fragen, warum Ereignis A nicht passiert ist, sondern warum Ereignis B stattdessen auftrat (Miller 2019).

Eine kontrafaktische Instanz ist eine minimal veränderte Dateninstanz, deren Vorhersage auf einen vordefinierten Wert geändert wurde. Sie kann neu generiert werden und muss nicht zwingend im Da-

tensatz vorhanden sein. (Molnar 2019)

Die Suche nach guten CF Instanzen ist eine Optimierungsaufgabe, mit zusammengesetzten, zu minimierenden Fehlerfunktionen, welche laut van Looveren & Klaise (2019) die folgenden Ziele adressieren:

- Die CF Vorhersage soll so nah wie möglich an der vordefinierten Vorhersage sein
- Die Veränderungen sollen im Vergleich zur Originalinstanz so gering wie möglich sein
- Das CF soll gut interpretierbar sein, indem es ähnlich der Trainingsdatenverteilung, besonders der Verteilung der kontrafaktischen Klasse, ist

*Implementierung, vgl. Tabelle A.7*

Für den XAIR wird die *Alibi* Implementierung *Counterfactuals guided by Prototypes*, nachfolgend CF-Proto genannt, ausgewählt. Diese basiert auf der Veröffentlichung von van Looveren & Klaise (2019). Zur Reduzierung des Zeitaufwands und zur Gewährleistung der Interpretierbarkeit reduziert diese die Unterschiedlichkeit des CFs zu dem jeweiligen Prototyp der vorgegebenen Klasse. Als Prototyp wird eine für eine Klasse repräsentative Dateninstanz bezeichnet. Ein weiterer Vorteil dieses Verfahrens gegenüber anderen ist seine Fähigkeit, die natürliche Ordnung kategorischer, ordinaler Features bei der Perturbation zu berücksichtigen, was ebenfalls die Realitätsnähe der CFs fördert. (van Looveren & Klaise 2019)

## Globale Methoden

Neben SHAP, das sowohl zu den lokalen als auch zu den globalen XAI-Methoden gezählt werden kann, wird die Permutation Feature Importance als globale Methode ausgewählt.

### Permutation Feature Importance (PFI)

Permutation Feature Importance (PFI) bietet global gesehen Aufschluss darüber, wie viel Einfluss ein Feature auf die Korrektheit der Vorhersage hat, d.h. wie sich der Vorhersagefehler des Modells nach einer Änderung des Feature-Wertes verändert. Dafür wird zunächst der Fehler des Modells durch eine beliebige Leistungsmetrik, bspw. die mittlere quadratische Abweichung (MSE, Mean Squared Error), gemessen. Anschließend wird jedes Feature permutiert, wobei sein Wert durch den einer anderen Dateninstanz ersetzt wird. Dies führt zu einem Bruch der Beziehung zwischen dem Feature und dem richtigen Vorhersageergebnis. Es wird der neue Fehler der permutierten Dateninstanz berechnet und die Wichtigkeit des Features durch die Differenz (oder durch das Verhältnis) des alten und neuen Fehlers bestimmt. (Molnar 2019)

*Implementierung, vgl. Tabelle A.6*

Die Implementierung von *ELI5* ist empfehlenswert, da sie eine Visualisierungsfunktion der Feature-Wichtigkeiten bietet. Allerdings ist bei der Verwendung von One-Hot Kodierung (OHE, One-Hot Encoding) zur Ermittlung der gesamten Feature-Wichtigkeit eine manuelle Aggregation der einzelnen, kodierten Features notwendig. Analog zu der Funktionsweise von OHE bei digitalen Schaltungen wird bei OHE im ML ein nominales Feature mit  $m$  Kategorien, durch  $m$  neue, binäre Features ersetzt, wobei ein Feature-Wert 1 („hot“) sein muss und alle anderen 0 (Harris & Harris 2010, S. 129).



Bei der alternativen ordinalen Kodierung wird jeder Kategorie eine Zahl zugewiesen. Die Zahlen haben allerdings keine sinnvolle Reihenfolge und sind nicht quantitativ zu verstehen (Han et al. 2012, S. 41). Ungewollt können sie allerdings vom Modell ebenfalls als geordnet interpretiert werden.

Wenn das Modell Teil einer Scikit-Learn Pipeline ist und somit die Datentransformationsschritte automatisch bei Eingabe einer Instanz in das Modell ausgeführt werden, ist zudem die Implementierung von *Scikit-learn* zu empfehlen. Bei One-Hot kodierten Features gibt sie automatisch die gebündelte Wichtigkeit der kodierten Features zurück.

Diese initiale Auswahl an XAI-Methoden soll durch einen modularen Aufbau des Empfehlungssystems nach Anforderung NFR4 einfach erweiterbar sein.

#### 4.1.3. Voraussetzungen für die Anwendung der XAI-Methoden

Nach der Auswahl der XAI-Methoden gilt es herauszufinden, welche Voraussetzungen für ihre Ausführung gegeben sein müssen. Sie werden anhand der Projektdokumentationen bzw. den wissenschaftlichen Veröffentlichungen ermittelt. Obwohl diese Rahmenbedingungen implementierungsspezifisch sind, werden diese als repräsentativ für die XAI-Methode angesehen.

Alle betrachteten Verfahren benötigen die Feature-Namen für die Beschriftungen und die Nachvollziehbarkeit der Ergebnisse. PFI benötigt für die Berechnung der Veränderung des Modellfehlers außerdem die dazugehörige Soll-Ausgabe (Label). Zudem benötigen alle Zugriff auf die Trainingsdaten; abhängig von der Methode variiert hingegen die benötigte Stichprobengröße. Außerdem muss unterschieden werden, ob sie *Vorbereitete Daten* oder *Engineered Features* für ihre Ausführung benötigt.

Vorbereitete Daten sind die in der gewünschten Granularität zusammengefassten, tabellarischen Daten, die innerhalb des Data Engineerings aus den Rohdaten durch Bereinigung invalider Features oder irrelevanter Dateninstanzen gewonnen werden. Als Engineered Features werden die auf das Modell und seine Aufgaben abgestimmten Features bezeichnet. Sie werden aus den Vorbereiteten Daten mithilfe der Vorverarbeitungsschritte (des Preprocessings) aufbereitet, z.B. skaliert, kodiert oder zu neuen Features aggregiert. (Google 2020a)

Die Schritte und Datenformatbezeichnungen des Preprocessings sind in Abbildung 4.1 veranschaulicht.

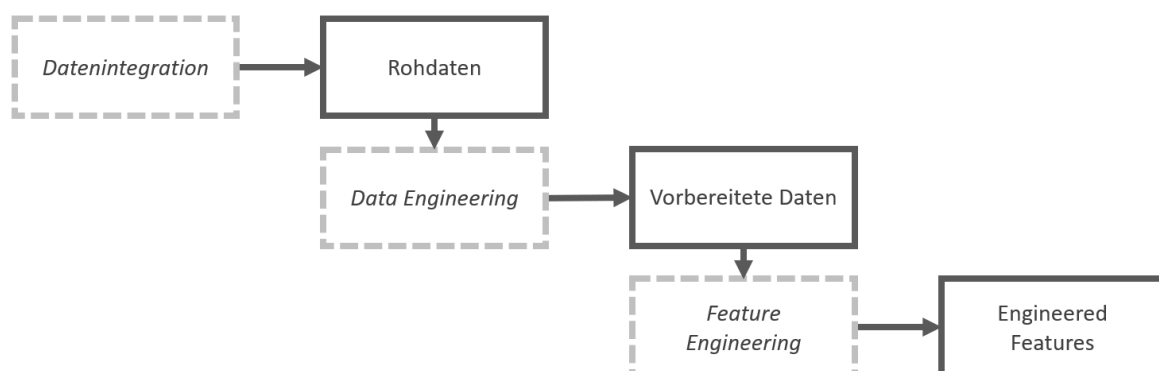


Abbildung 4.1.: Schritte und Formatbezeichnungen des Preprocessings, angelehnt an (Google 2020a)

CFProto und Anchors können nur bei Klassifikationsaufgaben verwendet werden. Für den Erhalt verständlicher Erklärungen ist außerdem der Zugriff auf die Preprocessing Operationen erforderlich. Die XAI-Methoden führen Datentransformationen vor dem Aufruf der eigentlichen Modellvorhersagefunktion aus. Die Anwendung der Methoden ohne Vorhandensein der Preprocessing Schritte wäre möglich, sofern alle kategorischen Features ausschließlich ordinal kodiert (nicht One-Hot kodiert) und alle numerischen Features unskaliert sind. Da numerische Daten standardmäßig normalisiert bzw. standardisiert werden, kommt dies in der Praxis kaum vor.

Eine weitere Bedingung für die Anwendung von CFProto ist außerdem die Rückgabe der Klassenwahrscheinlichkeiten der Vorhersagefunktion: Das Ergebnis einer Klassifikation mit  $n$  möglichen Klassen für eine Dateninstanz ist dabei ein Vektor mit  $n$  Wahrscheinlichkeitswerten.

Die identifizierten Voraussetzungen sind zusammengefasst in Tabelle 4.1 aufgeführt. Der Zugriff auf beide Formate der Trainingsdaten wird als gegeben gesehen, da das Training des Modells der Ausführung des XAIRs vorausgeht.

Tabelle 4.1.: Voraussetzungen für die Anwendung der XAI-Methoden

Voraussetzung		Methode					
		PDP + ICE	ALE	PFI	SHAP	Anchors	CFProto
Zugriff auf Modell Vorhersagefunktion			×	×			
		×			×	×	×
ML Aufgabe	Klassifikation	×	×	×	×	×	×
	Regression	×	×	×	×		
Rückgabe der Klassenwahrscheinlichkeiten							×
Zugriff auf Labels				×			
Zugriff auf Preprocessing Operationen						×	×
Datenformat	Vorbereitet Engineered Features	(*)	(*)	×	×	×	×
						(+)	(+)

(\*) Kodiert aber nicht skaliert, um Verfälschungen der Darstellungen numerischer Features zu vermeiden

(+) Da das Vorhandensein der Preprocessing Schritte obligatorisch ist, könnte man die benötigten vorbereiteten Daten aus den Engineered Features rekonstruieren

#### 4.1.4. Ableitung von Kriterien zur Beurteilung der Eignung von XAI-Methoden

Um die Informationen zu erhalten, die den XAIR für die Ermittlung der Güte der Methode braucht, muss zunächst die Frage geklärt werden, auf welche Faktoren eines spezifischen Modell-, Daten- und Nutzungskontexts bei der Beurteilung ihrer Anwendbarkeit geachtet werden muss. Wie in Kapitel 4.1.1 erörtert, zeichnen sich diese Faktoren dadurch aus, dass sie entweder die Anwendung einer XAI-Methode unmöglich machen, oder einen negativen, verfälschenden Einfluss auf die resultierende Erklärung oder ihre Interpretierbarkeit haben.

Nachfolgend wird genauer auf die anhand einer Literaturrecherche ermittelten Kriterien eingegangen.

## Korrelation der Daten

Bei allen initial ausgewählten XAI-Methoden handelt es sich um perturbationsbasierte Verfahren. Sie verändern die Eingabe-Features und gehen dabei, mit der Ausnahme von ALE, von einer Unabhängigkeit dieser untereinander aus. Wie bereits erwähnt ist bei korrelierenden Eingabe-Features die Erzeugung extrapolierte Datenpunkte möglich, die unrealistisch sein können. Diese werden ebenfalls für die Berechnung des Ergebnisses berücksichtigt und können es verfälschend beeinflussen.

Verallgemeinernd kann man daher sagen:

*„Bei einer hohen Korrelation des Datensatzes eignen sich perturbations- bzw. permutationsbasierte Verfahren eher weniger für eine Anwendung.“*

Im Gegensatz dazu eignet sich ALE bei einer hohen Korrelation des Datensatzes sehr gut, da die Feature-Werte nur lokal innerhalb eines kleinen Bereichs um den ursprünglichen Wert verändert werden, was die Extrapolation verhindert (Goldstein et al. 2015).

*„ALE eignet sich für die Anwendung bei stark korrelierenden Eingabe-Features (vergleichsweise) sehr gut.“*

## „Diskretisierbarkeit“ der Daten

Einige XAI-Methoden verwenden intern Diskretisierung in Form von Binning, d.h. sie teilen kontinuierliche, numerische Features in diskrete Teilmengen (Bins) auf.

Binning ist die einfachste Form der Diskretisierung, bei der die Bins vordefiniert sind und nicht auf die Verteilung angepasst werden (Garcia et al. 2013). Sie weisen entweder die gleiche Breite (Equal-Width Binning) oder die gleiche Frequenz an Datenpunkten (EqualFrequency) auf. Der Wert jeder Dateninstanz eines Bins wird dann durch den Bin-Mittelwert oder -Median ersetzt. (Han et al. 2012)

Die Auswahl eines guten Diskretisierungsverfahrens ist laut Garcia et al. (2013) schwierig und stark von den Daten abhängig. Die Verwendung eines auf die Daten angepassten Verfahrens ist bei keiner der Implementierungen möglich. Benutzerdefinierte Perzentile können dort zwar angegeben werden, allerdings werden diese dann für alle kontinuierlichen Variablen angewendet.

Bei Anchors fördert die Diskretisierung numerischer Features die Interpretierbarkeit der resultierenden „Wenn-Dann“-Regeln. Andernfalls wären diese sehr spezifisch und wiesen eine geringe Abdeckung des Feature-Eingaberaums auf. Allerdings kann das Binning bei sehr schiefen bzw. ungleichmäßig verteilten Daten, die dadurch eine schlechte „Diskretisierbarkeit“ aufweisen, die Entscheidungsgrenzen verschwimmen lassen, was die Güte der Methode mindert. (Molnar 2019)

Daher kann man sagen:

*„Wenn die Diskretisierbarkeit der Daten schlecht ist, dann ist Anchors eher weniger gut für die Anwendung geeignet.“*

Die Schiefe einer Feature-Verteilung, welche sich ebenfalls durch die Diskretisierbarkeit ermitteln lässt, beeinflusst außerdem die Interpretierbarkeit von ALE Plots. Da der Algorithmus den Feature-Wertebereich für die lokalen Perturbationen in EqualFrequency Bins aufteilt, führen schiefe Vertei-

lungen zu verzerrten, etwas schwerer interpretierbaren Visualisierungen (Molnar 2019):

*„Wenn die Diskretisierbarkeit der Daten schlecht ist, dann ist ALE unübersichtlicher und daher nur mittelmäßig gut für die Anwendung geeignet.“*

Die ausgewählte CFProto Implementierung verwendet intern ebenfalls die oben erwähnte Diskretisierung von Anchors (von *Alibi*), sofern die voreingestellten Parameter nicht angepasst werden.

Numerische Features werden diskretisiert, um die natürliche Ordnung von Features durch die Berechnung der Distanzen der einzelnen Kategorien zu ermitteln. Eine alternative Distanzmessung kann verwendet werden, die bei unabhängigen Kategorien ohne natürliche Ordnung bevorzugt wird (van Looveren & Klaise 2019). Die **Existenz ordinaler Features** kann demnach auch als Kriterium gesehen werden.

*„Wenn der Datensatz ordinale Features enthält und die Diskretisierbarkeit numerischer Features schlecht ist, dann ist CFProto eher weniger gut für die Anwendung geeignet.“*

## Präferenzen

### Leistung/Laufzeit der XAI-Methode

Wie in der Anforderungsanalyse in Anforderung FR4 aufgeführt, ist die Angabe der Präferenz der Leistung einer XAI-Methode gewünscht. Entscheidende Faktoren der Zeitkomplexität einer XAI-Methode sind die **Dauer des Modell-/Vorhersagefunktionsaufrufs**, die **Anzahl der Features** und die **Anzahl der betrachteten Dateninstanzen**. Letzteres ist bei der Ausführung der Methode durch den Nutzer regulierbar und wird daher nicht weiter betrachtet.

Je höher die Anzahl der Aufrufe des zu erklärenden Modells bzw. seiner Vorhersagefunktion ist, desto mehr fällt die dafür benötigte Zeit ins Gewicht. Daher kann eine XAI-Methode nach der erforderlichen **Anzahl der Aufrufe des Modells-/der Vorhersagefunktion** bewertet werden:

*„Je höher die Anzahl der Aufrufe des Modells bzw. der Vorhersagefunktion, desto schlechter sind die Leistungen der XAI-Methoden.“*

Perturbationsbasierte Verfahren brauchen bei einer höheren Anzahl an zu verändernden Features länger, weshalb für die Leistung außerdem gilt:

*„Je höher die Anzahl der Features eines Datensatzes, desto länger dauern perturbations- bzw. permutationsbasierte XAI-Methoden.“*

Andere sehr methodenspezifische Faktoren, die nicht allgemeingültig sind, bspw. der Einfluss der Zeitkomplexität der verwendeten MAB-Methode von Anchors, werden nicht betrachtet.

### Präferiertes Format der Erklärung

Auch die Angabe eines präferierten, spezifischen Erklärungsformats wurde genannt. Nach Sichten vorhandener XAI-Methoden wird diese Anforderung allerdings nur teilweise umgesetzt. Die Eingabe der Präferenz des Erklärungsumfangs soll weiterhin implementiert werden:

*„Wenn eine lokale oder globale Erklärung gewünscht ist und die XAI-Methode diese liefert, wird sie bevorzugt.“*

Allerdings wird die Angabe eines bevorzugten, spezifischen Erklärungsformates als nicht sinnvoll erachtet, weshalb die Umsetzung innerhalb des Projektumfangs nicht stattfindet. Dies wird damit begründet, dass es für einige Erklärungsformate sowohl generell, als auch innerhalb des prototypischen Empfehlungssystems, nur sehr wenige XAI-Methoden gibt. Daher wäre bei Forderung eines spezifischen Erklärungsformates eine leere Empfehlungsergebnismenge, oder eine, die keine Rücksicht auf die angegebene Präferenz nimmt, sehr wahrscheinlich.

### **Angabe der Zielgruppe/Komplexität der Erklärung**

Wie bei der Berücksichtigung eines spezifischen Erklärungsformates existieren auch für die Unterscheidung der Methodenkomplexität zu wenige XAI-Methoden. Daher entfällt auch die Umsetzung der Präferenz, eine Zielgruppe für die Förderung der Verständlichkeit der Methodenergebnisse anzugeben. Verständlichkeit ist außerdem subjektiv und schwer zu beurteilen.

### **Vorbereitungsaufwand**

Der Aufwand für die Vorbereitung, d.h. die Zeit der Einarbeitung in die XAI-Methode und in ihre Implementierung, hängt sehr von den Erfahrungen und Kenntnissen des Anwenders ab und ist daher sehr subjektiv. Er lässt sich nicht messen, nur aus Erfahrungen bewerten.

Im Hinblick auf die Weiterentwicklung des Empfehlungssystems bietet dieses Kriterium allerdings das Potential, Nutzern eine laut Ziegler & Loepp (2019) gewünschte, interaktive Einflussnahme auf die Bewertung der XAI-Methoden zu ermöglichen. Daher wird es in den Katalog der zu beachtenden Kriterien aufgenommen. Für die diesbezügliche Bewertung einer XAI-Methode kann man sagen:

*„Wenn ein geringer Vorbereitungsaufwand gewünscht ist, dieser für die Anwendung der XAI-Methode allerdings hoch ist, wird sie nicht bevorzugt.“*

*„Wenn ein hoher Vorbereitungsaufwand akzeptiert oder egal ist, eignen sich auch XAI-Methoden mit hohem Vorbereitungsaufwand.“*

### **Datenspezifische Eigenschaften kritischer Features**

Abschließend werden diese identifizierten Kriterien außerdem um die Kriterien der Korrelation und der Diskretisierbarkeit der „Features of Interest“ (FOI) eines Datensatzes ergänzt, sofern solche vorhanden sind. Als FOI werden von nun an Features bezeichnet, die eine menschliche Voreingenommenheit widerspiegeln und sich negativ auf die Modellentscheidung auswirken können. Sie bieten das Potential zur Diskriminierung und bedürfen daher besonderer Aufmerksamkeit. Gängige, im Rahmen des Allgemeinen Gleichbehandlungsgesetzes (AGG) in §1 verbotene Beispiele hierfür sind: Rasse oder ethnische Herkunft, Geschlecht, Religion oder Weltanschauung, Behinderung, Alter oder sexuelle Identität (Antidiskriminierungsstelle des Bundes 2006).

Das ausführliche Examinieren dieser Features durch bspw. eine visuelle Darstellung mittels PDP/ICE oder ALE Plots ist zur Vermeidung diskriminierender Modelle sinnvoll. FOI, die stark mit den anderen Eingabe-Features korrelieren, deuten außerdem auf Proxy-Features hin. Proxy Features wirken unverdächtig und neutral, können allerdings durch eine hohe Korrelation mit dem FOI oder mit dem Vorhersageergebnis die Mitglieder einer geschützten Gruppe unverhältnismäßig stark benachteiligen

(Prince & Schwarcz 2020). Um solches Diskriminierungspotenzial zu vermeiden ist die Ausgabe weiterer Hinweise und somit auch die Erhebung ihrer Ausprägungen vorgesehen.

Zusammenfassend wird die Eignung der XAI-Methoden hinsichtlich der in Abbildung 4.2 aufgeführten Kriterien des Modell-, Daten- und Nutzungskontexts beurteilt.

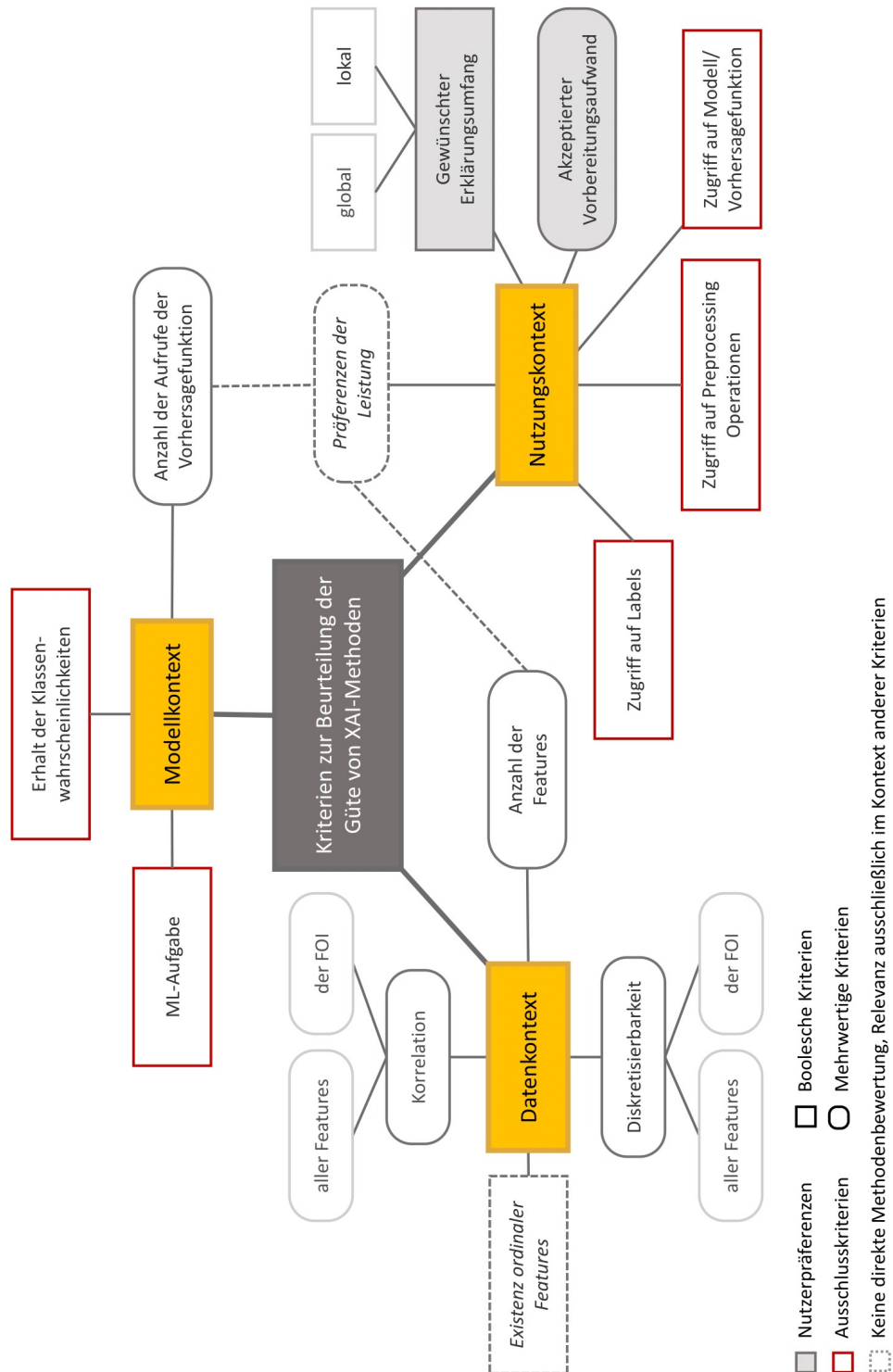


Abbildung 4.2.: Übersicht der für die Beurteilung der Eignung einer XAI-Methode identifizierten Kriterien des Modell-, Daten- und Nutzungskontexts

## 4.2. Auswahl einer Wissensrepräsentation

Nach der Ableitung der für die Empfehlungsgenerierung relevanten Kriterien muss nun eine Repräsentation des Wissens gewählt werden, mit der ihre Auswirkungen auf die Eignung der XAI-Methoden formalisiert werden kann. Dafür werden zuerst die Herausforderungen bei der Erhebung und Beurteilung der Kriterienwerte aufgeführt. Diese berücksichtigend, wird anschließend eine Wissensrepräsentation gewählt.

### 4.2.1. Problem der Erhebung der Kriterienwerte und ihrer Auswirkungen

Bei der Erhebung der in Kapitel 4.1.4 identifizierten Kriterien treten einige Probleme auf, die am Beispiel der Korrelation nachfolgend erläutert werden.

Gängige Korrelationsmessungen erfolgen paarweise durch Berechnung eines Korrelationskoeffizienten, der die Stärke des statistischen Zusammenhangs zwischen zwei Variablen quantifiziert (Fahrmeir 2004, S. 148). Wie kann man jedoch die Stärke der Korrelation eines kompletten Datensatzes messen? Ausgehend von einer paarweisen Berechnung der Korrelationskoeffizienten stellen sich folgende Fragen:

- Ab welcher Korrelationsstärke gelten zwei Features als korrelierend?
- Ab wann bzw. ab welcher Anzahl an hohen, paarweisen Korrelationen pro Feature gilt ein Feature als korrelierend bzgl. des gesamten Datensatzes?
- Wann bzw. ab welcher Anzahl an korrelierenden Features gilt ein ganzer Datensatz als korrelierend?

Es existieren keine Schwellwerte, die die Beantwortung der oben aufgeführten Fragen ermöglichen.

Das gleiche Problem fehlender Schwellwerte tritt auch bei den anderen, nicht-zweiwertigen Kriterien, z.B. der Diskretisierbarkeit, auf. Abgesehen von der Frage, was man als „Maß für die Anwendbarkeit von Diskretisierungsverfahren“ ansehen kann, fehlen Grenzen, ab wann ein Feature, und ab welcher prozentualen Anzahl ein ganzer Datensatz, als gut bzw. schlecht diskretisierbar gilt.

Zudem gibt die Literatur nur in vagen Begriffen Auskunft über die Nachteile bei der Anwendung von XAI-Methoden, die durch die identifizierten Kriterien entstehen. Eine genaue Eignungsminde- rung ist nicht metrisch messbar. Man kann lediglich sagen, „*SHAP ist für stark korrelierende Daten eher ungeeignet*“, nicht aber: „*SHAP weist eine Eignung von  $p\%$  auf, wenn der Anteil der Features, bei denen die paarweise Korrelation von mindestens  $a$  Features mindestens  $b$  groß ist, größer als  $c$  ist*“.

Eine genaue Methodenbeurteilung hinsichtlich der sehr subjektiven Vorbereitungszeit und ihrer Leistung, die sich aus der Feature-Anzahl und der Zugriffszeit auf das zu erklärende Modell bzw. seiner Vorhersagefunktion ergibt, ist ebenfalls unmöglich. Eine relative Einschätzung der beiden leistungsbeeinflussenden Parameter ist aufgrund der Unterschiedlichkeit der Datensätze und Modelle und einem Mangel an Referenzwerten unmöglich. Abgesehen davon sind Zugriffszeiten von physischen Faktoren wie bspw. der Rechenleistung der Hardware oder den Netzwerkverbindungen abhängig. Diese sind nicht-deterministisch und können nicht ermittelt werden.

Bei der Auswahl der Wissensrepräsentation und der Inferenz des Expertensystems muss daher das Problem der Ungenauigkeit adressiert werden.

#### 4.2.2. Wahl der zu verwendenden Logik und Wissensrepräsentation

Mithilfe der klassischen, zweiwertigen Logik lassen sich die oben genannten, vagen Aussagen nicht formulieren. Computer, die nach dieser agieren und daher eindeutige (crisp) Werte, 0 und 1, benötigen, können mit nicht eindeutig zuweisbaren (fuzzy) Begriffen wie „mittelstark korrelierend“ oder „stark korrelierend“ nicht umgehen. Die Fuzzy-Logik bietet eine Lösung dieser Probleme.

Durch die Einführung mehrwertiger Wahrheitswerte können Aussagen wie „der Datensatz ist stark korrelierend“ approximativ mit einem Wahrheitswert zwischen 0 und 1 bewertet werden. So wird eine Definition von Parameterschwellwerten vermieden. Zudem wird der Volatilität des Systems entgegengewirkt (NFR6), indem kleine Änderungen der Eingaben durch Überschreiten eines intern gesetzten Grenzwerts keine großen Änderungen des Empfehlungsergebnisses verursachen können. Durch die Möglichkeit einer ungefähren Angabe wird der Nutzer von der tiefen, datenanalytischen Auseinandersetzung mit den Trainingsdaten verschont. Neben der Steigerung der Benutzerfreundlichkeit (NFR3) reduziert das, wie von der Nutzerzielgruppe gewünscht, den Aufwand der Parametereingabe (NFR1) und senkt dadurch die Hemmschwelle für die Anwendung des XAIRs. Er kann somit explorativ befragt werden, auch wenn der dem Modell zugrundeliegende Datensatz nicht ausführlich examiniert wurde. Außerdem wird die einfache und intuitive Formulierung der in FR4 gewünschten Berücksichtigung nutzerspezifischer Präferenzen ermöglicht.

Die Fuzzy-Logik löst zudem das Problem der vagen Bewertungen der XAI-Methodeneignungen, indem sie eine qualitative Modellierung des ungenauen, in der Literatur zu findenden Expertenwissens ermöglicht. Als Logik des approximativen Schließens erlaubt sie außerdem eine Vagheit und Nicht-Eindeutigkeit möglicher Empfehlungsergebnisse (Zadeh 1975) und spiegelt somit die Realität wider.

Aus den oben genannten Gründen wird der entstehende XAIR mithilfe eines Fuzzy-Expertensystems umgesetzt. Auf die Konfiguration und die Umsetzung der System-Komponenten wird im nachfolgenden Kapitel eingegangen.

### 4.3. Konfiguration des Fuzzy-Expertensystems

Für die Implementierung eines Fuzzy- Expertensystems müssen zuerst seine Ein- und Ausgabeveriablen definiert werden. Im Zuge dessen wird geklärt, wie die Eingabewerte (automatisiert) ermittelt werden können. Anschließend folgt die Konfiguration der in Kapitel 2.4.2 aufgeführten Komponenten eines Fuzzy- Expertensystems, der Fuzzifizierung, der Fuzzy-Inferenz und der Defuzzifizierung.

#### 4.3.1. Definition der Ein- und Ausgaben

In Kapitel 4.1.4 wurden die Voraussetzungen und Beurteilungskriterien der XAI-Methoden identifiziert. Für diese muss nun jeweils ein Eingabeparameter definiert werden, sodass der XAIR sie für die Empfehlungsgenerierung in Betracht ziehen kann.

Die Voraussetzungen (Tabelle 4.1, S. 33) sind binär und können auf weniger Eingabeparameter des



Systems reduziert werden. Ihre Erfüllung ist für die Anwendung einer XAI-Methode obligatorisch, weshalb sie als Ausschlusskriterien bzgl. ihrer Empfehlung angesehen werden.

Obwohl die Kriterien „Globale Erklärung“ und „Lokale Erklärung“ ebenfalls boolesch sind, werden sie nicht als Ausschlusskriterien gesehen. Sie fließen als Nutzerpräferenz (Anforderung FR4) in die Empfehlungsentscheidung mit ein und beeinflussen so zwar die Eignung von XAI-Methoden, schließen aber keine Methoden aus.

Die datenbezogenen Parameter, die Korrelation (der FOI) und Diskretisierbarkeit (der FOI), können fuzzy eingeschätzt, aber auch exakt durch eine Datenanalyse ermittelt und daher crisp angegeben werden. Alle weiteren eignungsreduzierenden Eingaben werden als fuzzy Einschätzung des Nutzers entgegengenommen. Aufgrund der Unterschiedlichkeit möglicher Werte und einem Mangel an Referenzen ist eine exakte Angabe durch den Nutzer nicht möglich.

Die nachfolgende Liste führt alle Eingabeparameter des XAIRs auf:

- **Verfügbarkeit des Modells** (bool, Ausschlusskriterium)
- **Klassifikationsaufgabe** (bool, Ausschlusskriterium)
- **Erhalt der Klassenwahrscheinlichkeiten** (bool, Ausschlusskriterium)
- **Zugriff auf Labels** (bool, Ausschlusskriterium)
- **Zugriff auf Preprocessing Operationen** (bool, Ausschlusskriterium)
- FOI (Liste mit Feature Namen)
- Korrelation (crisp/fuzzy)
- Korrelation der FOI (crisp/fuzzy)
- Diskretisierbarkeit (crisp/fuzzy)
- Diskretisierbarkeit der FOI (crisp/fuzzy)
- Präferenzen der Leistung (fuzzy)
- Anzahl der Features (fuzzy)
- Zugriffszeit Modell/Vorhersagefunktion (fuzzy)
- Vorbereitungsaufwand (fuzzy)
- Präferenz einer globalen Erklärung (bool)
- Präferenz einer lokalen Erklärung (bool)
- Vorhandensein ordinaler Features (bool)

Formulierungen für die Beschriftungen für die Eingabe sind in Anhang A.3 zu finden. Diese sind besonders bei der Integration neuer XAI-Methoden (bzgl. NFR4)) in das System für das korrekte Setzen der Ausschlusskriterien relevant.

Die Ausgabevariablen des Fuzzy-Expertensystems sind die für den XAIR in Kapitel 4.1.2 ausgewählten XAI-Methoden:

- Partial Dependence Plots und Individual Conditional Expectation (PDP + ICE)

- Accumulated Local Effects (ALE)
- Permutation Feature Importance (PFI)
- SHapley Additive exPlanations (SHAP)
- Anchors (Anchors)
- Counterfactuals guided by Prototypes (CFProto)

#### 4.3.2. Erhebung und Ermittlung der Systemeingaben

Bei einer eigenständigen Verwendung des XAIRs werden die Eingabeparameter über eine Weboberfläche vom Nutzer entgegengenommen (FR3). Wie bereits erwähnt können datenbezogene Parameter der Korrelation (der FOI) und Diskretisierbarkeit (der FOI) entweder nach einer manuell durchgeführten Datenanalyse crisp, oder nach eigener Einschätzung fuzzy sein.

Nach NFR1 soll das System in eine ML-Pipeline integrierbar sein. Im Hinblick auf diese Verwendung ist eine automatisierte Ermittlung der für eine Ergebniserzeugung benötigten Dateneigenschaften notwendig. Alle nicht datenbezogenen Eingabeparameter müssen dafür in der Pipelinekonfiguration hinterlegt sein, da sie nicht in einer Datenanalyse bestimmt werden können.

Das in Abbildung 4.3 zu sehende Use Case Diagramm stellt dar, wie der Nutzer sich eine Empfehlung vom XAIR als Web-Anwendung, und automatisiert innerhalb einer ML-Pipeline einholen kann.

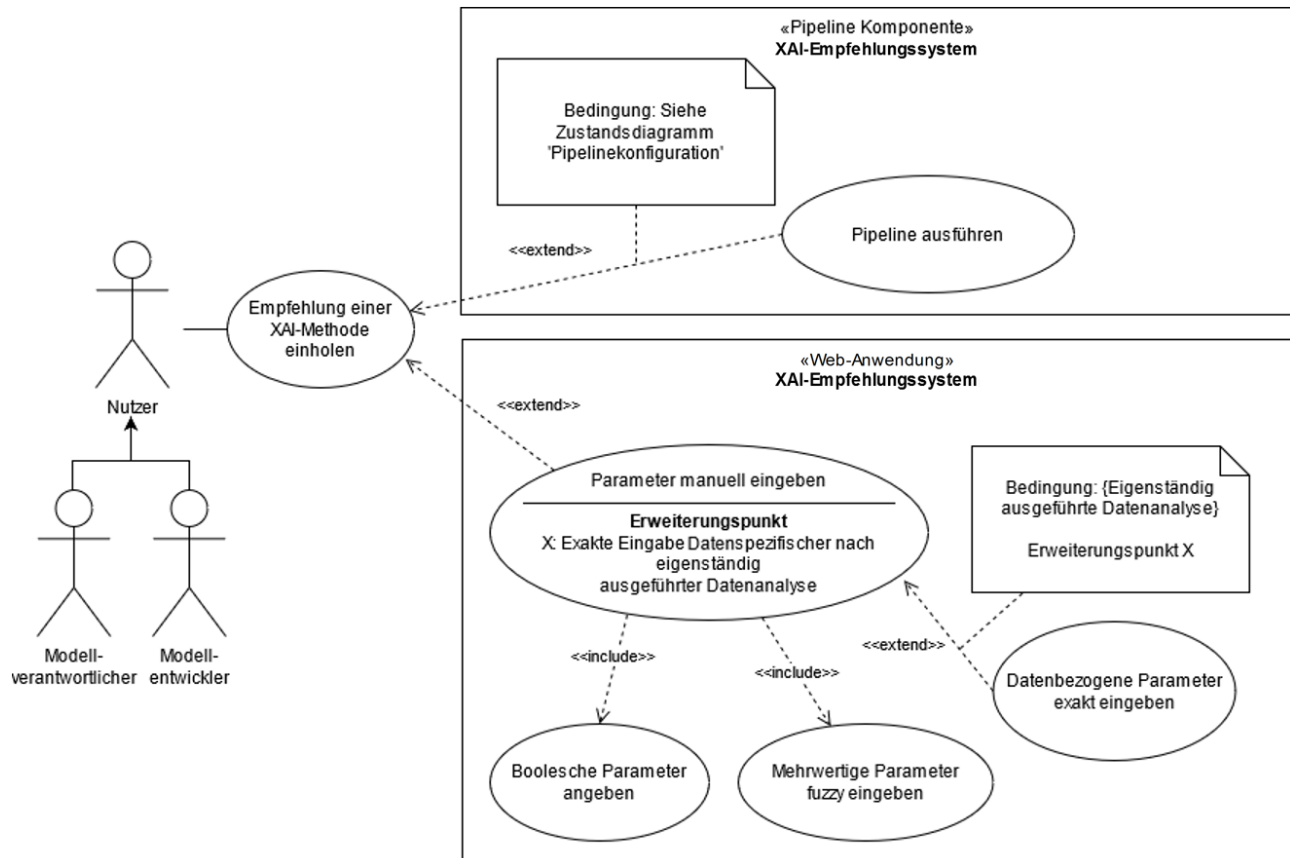


Abbildung 4.3.: Use Case Diagramm des Empfehlungserhalts

Der Prozess der Pipelinekonfiguration, welche für die Verwendung des Empfehlungssystems innerhalb einer ML-Pipeline notwendig ist, ist im Zustandsdiagramm in Abbildung 4.4 dargestellt.

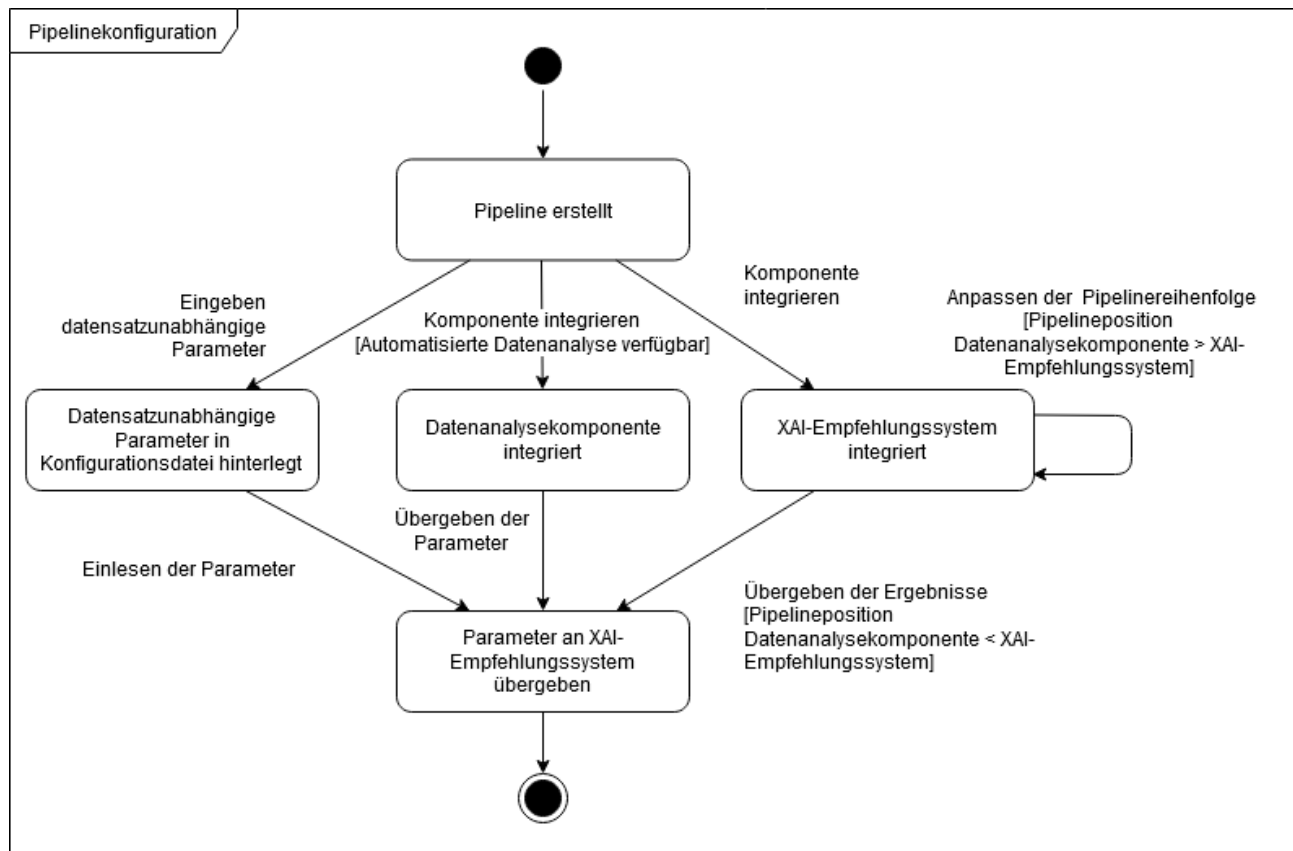


Abbildung 4.4.: Zustandsdiagramm der Konfiguration einer ML-Pipeline

Das Feedback während der iterativen Evaluation des Artefakts gab Aufschluss, dass Nutzern die Bedeutung und Messung der Korrelation für eine Eingabe in den XAIR nicht intuitiv klar ist (vgl. Kapitel 4.2.1).

Um dem Nutzer Hilfe bei der Einschätzung des Datensatzes zu geben, wird nachfolgend näher auf die Ermittlung der datenbezogenen Eingabeparameter eingegangen. Die nun aufgeführten Ansätze dienen im Hinblick auf eine Erweiterung des Nutzungsszenarios auf den Einsatz in einer ML-Pipeline (NFR1) als Konzept für eine mögliche, dem XAIR vorangehende Datenanalyse-Komponente. Diese soll automatisiert die Eingabeparameter der Korrelation und der Diskretisierbarkeit (der FOI) aus den Daten gewinnen.

Bevor eine Strategie zur Ermittlung der Werte erstellt werden kann, stellt sich die Frage, anhand welcher Daten die Analyse erfolgt. Es stehen die rohen, die vorbereiteten oder die für das Modell verwertbaren Daten (Engineered Features) zur Verfügung.

### Identifikation der für die Analyse geeigneten Daten

Für eine Analyse wäre angemessen, die von der XAI-Methode empfangenen Daten zu betrachten, da auf ihnen die Perturbationen bzw. das Binning durchgeführt werden. Wie in Tabelle 4.1 (S. 33) aufgeführt, ist das geforderte Format allerdings methodenspezifisch und unterschiedlich.

Die Betrachtung der Engineered Features ist sinnvoll, da die resultierende Modellvorhersage auf diesen basiert. Allerdings sind sie durch die Kodierung und Skalierung nicht wirklich menschenverständlich. Eine Interpretierbarkeit der Ergebnisse der Datenanalyse ist jedoch anzustreben, um Nachvollziehbarkeit der Eingabeparameter zu gewährleisten. Sie kann außerdem einer erneuten Überprüfung und Qualitätssicherung des Preprocessings dienen. Lediglich die rohen und die vorbereiteten Daten sind für den Nutzer aufschlussgebend, wobei Rohdaten aufgrund ihrer unstrukturierten Form nicht analysiert werden können.

Im Zuge dieser Überlegungen wird eine Empfehlung für die Ausführung des Preprocessings gegeben, die in Anhang A.4 zu finden ist. In dieser wird eine Aufteilung des Preprocessings in zwei Teile vorgeschlagen: Zuerst werden die non-destruktiven Operationen ausgeführt, bei denen die ursprüngliche Repräsentation des Features (vor der Transformation) wiederherstellbar ist. Anschließend folgen die destruktiven Operationen, die nicht bijektiv und daher nicht rückgängig machbar sind. Diese Reihenfolge fördert die Transparenz des Datenvorbereitungsprozesses und beugt somit eventuellen grundlegenden Veränderungen oder Verfälschungen der Daten vor.

Für die Eignungsprüfung der XAI-Methoden werden die entwickelten Empfehlungen aus Anhang A.4 berücksichtigt. Die Datenanalyse findet daher auf den Engineered Features statt, die zwar den destruktiven Preprocessing Operationen unterzogen wurden, allerdings nicht kodiert und unskaliert sind.

### Berechnung datenspezifischer Eingabeparameter

#### Korrelation

Die Problematik bei der Ermittlung einer Messung der Gesamtkorrelation des Datensatzes wurde in Kapitel 4.2.1 bereits angeschnitten. Obwohl die Fuzzy-Logik das Ausdrücken von Ungenauigkeiten erlaubt, besteht bei der Erhebung der Gesamtkorrelation weiterhin das Problem, dass gängige Korrelationskoeffizienten ausschließlich für die Berechnung der paarweisen Korrelation zweier Features desselben Datentyps (numerisch, nominal, ordinal) verwendet werden können.

Für quantitative Features (kontinuierliche, reellwertige Variablen) ist der Korrelationskoeffizient  $\rho$  nach Pearson, der die Stärke und Richtung einer linearen Abhängigkeit misst, der de-facto Standard (Baak et al. 2020, S. 3). Korrelationen ordinaler Variablen können unter anderem durch den Rangkorrelationskoeffizienten nach Spearman ermittelt werden, der den Pearson Korrelationskoeffizient auf die Ränge der ordinalen Variablen anwendet (Fahrmeir 2004, S. 142). Dieser geht allerdings von gleichmäßigen Abständen zwischen den Variablenwerten aus (Baak et al. 2020, S. 5).

Zusammenhänge nominaler Features ohne natürliche Ordnung können durch verschiedene Assoziationsmaße, die auf der Betrachtung der gemeinsamen oder bedingten Häufigkeitsverteilung in Kontingenztafeln basieren, ermittelt werden. Beispiele dafür sind der  $\chi^2$  Koeffizient und der (korrigierte)

Kontingenzkoeffizient. (Fahrmeir 2004, S. 168)

Der für nominale Variablen anwendbare Korrelationskoeffizient Cramér's  $\Phi$  kann auch für ordinale und diskretisierte Intervallvariablen angewendet werden. Allerdings sind seine Ergebnisse stark von dem dafür ausgewählten Diskretisierungsverfahren abhängig. (Baak et al. 2020, S. 6)

Eine automatisierte Korrelationsberechnung mit nicht auf die Feature-Verteilung angepasster Diskretisierung wird daher nicht in Erwägung gezogen.

Da ein typischer Datensatz eine Kombination mehrerer Datentypen enthält und die Korrelation aufgrund der unscharfen XAI-Methodeneinschätzung nicht nur für zwei, sondern über alle Features berechnet werden soll, kommt keines der zuvor erwähnten Berechnungsmaße infrage.

Abhilfe bietet der globale  $\Phi_K$  Korrelationskoeffizient  $g_k$ , der Auskunft darüber gibt, wie gut jedes Feature aus den anderen gegebenen modelliert werden kann (Baak et al. 2020, S. 5). Er gibt datentypunabhängig die stärkste Korrelation einer Variablen  $k$  und einer Linearkombination aller anderer Variablen, mit einer Zahl zwischen 0 und 1, an.

Die Ermittlung erfolgt aus der Kovarianzmatrix  $V$  (Baak et al. 2020, S. 5):

$$g_k = \sqrt{1 - [V_{kk} * (V^{-1})_{kk}]^{-1}} \quad (8)$$

Dafür werden die Varianzen der Intervallvariablen auf 1 skaliert, die kategoriieller Features als „undefined“ angenommen und die  $\Phi_K$  Korrelationsmatrix  $C$  für  $V$  verwendet.  $\Phi_K$  verwendet zur Berechnung von  $C$  intern Binning in Dezile und setzt  $V = C$ . (Baak et al. 2020, S. 13)

Um eine Anwendbarkeit bei schlecht diskretisierbaren Features sicherzustellen, wurde ein Vergleich mit Pearson's  $\rho$  vorgenommen, der auf Intervallvariablen ausgelegt ist. Dabei wurden die Ergebnisabweichungen der beiden Verfahren mit sehr ungleichen Feature-Verteilungen ermittelt. Es wurden diverse Verteilungen mit 1000 bzw. 10000 Dateninstanzen erzeugt und die Differenz der paarweisen Korrelationen beider Verfahren in 50 bzw. 500 Iterationen gemessen. Die Ergebnisse sind in Anhang A.5 aufgeführt und ergeben, dass die durchschnittlichen Unterschiede der Korrelation einer Feature-Kombination mit maximal rund 0.0694 bzw. 0.0285 marginal sind.

Bei einer bivariaten Normalverteilung entspricht  $\Phi_K$  außerdem Pearson's  $\rho$ . Zudem besitzt  $\Phi_K$  eine Korrektur des statistischen Rauschens. Im Gegensatz zu Cramér's  $\Phi$  ist er stabil gegenüber der Anzahl der für eine Diskretisierung gewählten Bins pro Intervallvariable und daher eindeutig interpretierbar. (Baak et al. 2020)

Die Anwendung des  $g_k$  liefert die globale Korrelation jedes Eingabe-Features bzgl. aller anderen. Eine beispielhafte grafische Ausgabe des  $g_k$  anhand des UCI Adult Datensatzes (Dua & Graff 2017) ist in Abbildung 4.5 zu sehen.

Diese vielen Korrelationswerte müssen zu einem globalen zusammengefasst werden, der die Gesamtkorrelation des Datensatzes repräsentiert und dem Fuzzy-Expertensystem als Eingabe dient. Bei diesem soll die Anzahl und Stärke starker Korrelationen mehr ins Gewicht fallen als die exakte Ausprägung schwacher, da die Güte der XAI-Methoden hauptsächlich durch starke Korrelationen beeinflusst wird und sich schwache neutral auf sie auswirken.

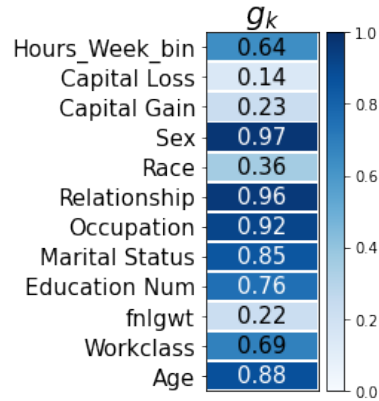


Abbildung 4.5.: Grafische Ausgabe des  $g_k$  am Beispiel der Features des UCI Adult Datensatzes

Die Ermittlung einer Aggregationsmethode wurde in mehreren Iterationen vorgenommen, deren Variationen nachfolgend aufgeführt werden.

*„Je höher der Durchschnitt der einzelnen Korrelationswerte, desto stärker ist die Gesamtkorrelation des Datensatzes“*

Die einfache Verwendung des Durchschnitts ist für die Darstellung der Gesamtkorrelation nicht aussagekräftig, da hohe Korrelationswerte von niedrigen ausgeglichen werden.

*„Je höher die Mehrheit der Korrelationswerte, desto stärker ist die Gesamtkorrelation des Datensatzes“*

Für eine Abschätzung der globalen Korrelation ist die Betrachtung der Verteilung der einzelnen Korrelationswerte möglich. Allerdings ist das Lagemaß der Schiefe durch einen Ausgleich hoher und niedriger Werte nicht aufschlussgebend.

*„Je stärker die maximale Korrelation einer signifikanten Anzahl der Features (z.B. 75%), desto stärker ist die Gesamtkorrelation des Datensatzes“*

Um das Problem der Mittelung hoher und niedriger Korrelationswerte zu adressieren, wird die Einordnung der Gesamtkorrelation durch eine Kombination der Anzahl und der Höhe der Korrelationswerte überprüft. Dafür kann das Lagemaß des Perzentils verwendet werden, welches resistent gegenüber Ausreißern ist. Z.B. kann der Wert des oberen Quantils (75%-Perzentil) als Gesamtkorrelationswert in Betracht gezogen werden. Das gewährleistet, dass die Mehrheit der Features geringer korrelieren und kann daher als repräsentativ für den Datensatz gesehen werden. Allerdings müssen dafür erneut Annahmen über einen Schwellwert, wie viele Features stark korrelieren müssen, getroffen werden, sodass der Datensatz als stark korrelierend gilt. Daher wird der Ansatz verworfen.

*„Je stärker die durchschnittliche Gesamtkorrelation der Features oder je höher die Anzahl stark korrelierender Features, desto stärker ist die Gesamtkorrelation des Datensatzes“*

In dieser Iteration werden sowohl die durchschnittliche Stärke der Korrelation der einzelnen Features, als auch der Anteil stark korrelierender Features für die Berechnung der Gesamtkorrelation berücksichtigt.

Um den Anteil hoher Korrelationen abzuschätzen, ordnet man zunächst die einzelnen Korrelationwerte Fuzzy-Mengen zu, die gängigen Daumenregeln der Korrelationsinterpretationen entsprechen. Diese sind in den Tabellen 4.2 dargestellt.

Tabelle 4.2.: Daumenregeln der Interpretation von Korrelationskoeffizienten, bspw. nach (Evans 1996, Hinkle et al. 2003)

Korrelation	Interpretation	Korrelation	Interpretation
0.00 – 0.19	Sehr schwach	0.00 – 0.30	Sehr schwach
0.20 – 0.39	Schwach	0.30 – 0.50	Schwach
0.40 – 0.59	Mittel	0.50 – 0.70	Mittel
0.60 – 0.79	Stark	0.70 – 0.90	Stark
0.80 – 1.00	Sehr stark	0.90 – 1.00	Sehr stark

Anschließend wird der Anteil der Zugehörigkeiten aller Elemente zu den Fuzzy-Mengen STARK und SEHR STARK relativ zu der Summe aller Zugehörigkeitswerte aller Elemente gesehen. Sollte der Anteil nicht so groß sein, ist die potenziell mittelstarke Korrelation der anderen Features, die als wenig korrelierend eingeschätzt sind, nicht zu vernachlässigen. Daher wird auch die Durchschnittskorrelation berechnet, die das „Grundrauschen“ der Korrelationen innerhalb des Datensatzes widerspiegelt.

Der finale Gesamtkorrelationswert kann mithilfe des fuzzy-logischen OR-Operators (Gleichung 7) gebildet werden, welcher den maximalen Fuzzy-Wert der beiden Variablen „Anteil hoher Korrelationen“ und „Durchschnittliche Korrelation“ wählt. Tabelle 4.3 zeigt diese fuzzy OR-Verknüpfung, wobei bei dieser zur Bewahrung der Übersichtlichkeit die Variablen nur drei statt fünf Zugehörigkeitsfunktionen besitzen. Unrealistische Parameterkombinationen sind grau hinterlegt.

Tabelle 4.3.: Bildung des Gesamtkorrelationswertes durch die fuzzy OR-Verknüpfung

Anteil hoher Korrelationen	Durchschnittliche Korrelation	Gesamtkorrelation des Datensatzes
L	L	L
L	M	M
L	H	H
M	L	M
M	M	M
M	H	H
H	L	H
H	M	H

Obwohl die Ermittlung des für eine Eingabe geeigneten Korrelationswerts auf diese Weise bei diversen getesteten Datensätzen, unter anderem dem UCI Adult Datensatz, schlüssig ist, muss dafür zweifach fuzzifiziert und ein eigenes Regelwerk implementiert werden. Alternativ liefert die Betrachtung der Durchschnittskorrelation und die Streuung der einzelnen globalen Werte vergleichbare Ergebnisse mit weniger Berechnungsaufwand.

Die Verwendung der empirischen Standardabweichung als Maß für die Streubreite der gemessenen

Korrelationswerte beschreibt die durchschnittliche Entfernung aller einzelnen Feature-Korrelationen  $(x_1, x_2, \dots, x_n)$  zum Durchschnittskorrelationswert  $(\bar{x})$ . Sie wird durch folgende Formel beschrieben (Han et al. 2012, S. 51):

$$s = + \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (9)$$

Neben der Standardabweichung wurde auch die Verwendung der Mittleren Absoluten Abweichung (MAD, Mean Absolute Deviation) in Betracht gezogen (Han et al. 2012, S. 114):

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| \quad (10)$$

Dieser misst den durchschnittlichen (absoluten) Abstand zum Durchschnitt und ist durch das Fehlen der Quadrierung im Gegensatz zu der Standardabweichung resistenter gegenüber Ausreißern (Han et al. 2012, S. 115). Durch die Begrenzung des Wertebereichs der Korrelation auf  $[0, 1]$  ist dies allerdings nicht von großer Bedeutung.

Da die Unterschiede der Ergebnisse minimal sind, wird der MAD für die Berechnung der Korrelation gewählt und für einen Erhalt des Gesamtkorrelationswerts mit dem Korrelationsdurchschnitt kombiniert. Dabei wird der MAD halbiert, was zu denselben Fuzzy-Ergebnissen der Gesamtkorrelationsstärke der in Tabelle 4.3 aufgeführten Aggregationen führt:

$$corr = \bar{x} + \frac{1}{2} MAD \quad (11)$$

Im Gegensatz zur vorigen Iteration wird nun nicht mehr der Anteil (sehr) starker Korrelationen, sondern die „erwartete/durchschnittliche Stärke starker Korrelationen“ berücksichtigt:

*„Je stärker die zu erwartenden Korrelationen stark korrelierender Features des Datensatzes, desto stärker die Gesamtkorrelation des Datensatzes“.*

Der resultierende und fuzzifizierte Korrelationswert des Gesamtdatensatzes ist deckungsgleich mit menschlichen, subjektiven Einschätzungen. Zur Nachvollziehbarkeit dieser Aussage sind in Anhang A.6 Korrelationsberechnungen einiger Beispieldatensätze aufgeführt.

### Diskretisierbarkeit

Wie bereits erwähnt teilen einige der ausgewählten XAI-Methoden kontinuierliche Features in diskrete Bins auf und sind daher nicht so gut geeignet, wenn Features nicht gut diskretisierbar sind. Zur Ermittlung der Eingabe für den XAIR stellt sich die Frage, was man als „Maß für die Anwendbarkeit von Diskretisierungsverfahren“ ansehen kann.

Die Betrachtung der Größe des Wertebereichs allein ist nicht sinnvoll, denn sollten die Datenpunkte gleichverteilt sein, stellt das Binning kein Problem dar. Obwohl die Verteilung der Werte eine Rolle spielt, ist eine Einschätzung der Diskretisierbarkeit über eine Analyse der Lagemaße, bspw. der Schiefe, durch einen bereits erwähnten Ausgleich von hohen und niedrigen Werten auch nicht möglich.



Bei der Verwendung von EqualWidth Binning fällt auf: Je ungleicher die Feature-Verteilung, desto stärker unterscheiden sich die Bins in der Anzahl der in ihnen liegenden Datenpunkte, und desto ungeeigneter scheint diese Art der Aufteilung. Dies ist in Abbildung 4.6 veranschaulicht, welche Histogramme der Features „Age“ und „Capital Gain“ des UCI Adult Datensatzes zeigt. Wie ersichtlich fallen beim Feature „Capital Gain“, im Gegensatz zum nur leicht rechtsschiefen Feature des Alters, fast alle Dateninstanzen in einen Bin. Dies deutet auf eine sehr schlechte Diskretisierbarkeit hin.

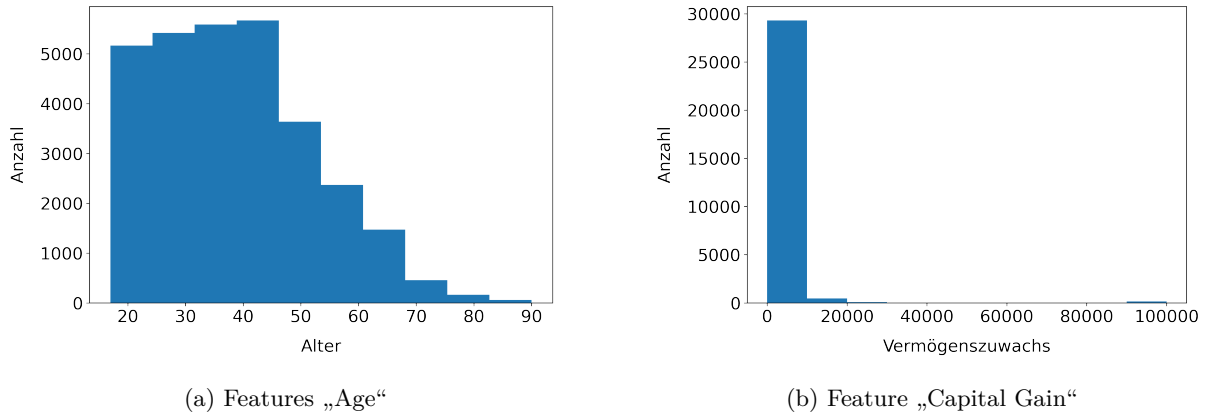


Abbildung 4.6.: EqualWidth Binning der unterschiedlich gut diskretisierbaren Features „Age“ und „Capital Gain“ des UCI Adult Datensatzes

Analog kann man für EqualFrequency Binning sagen, dass die Anwendung für eine Verteilung nur sinnvoll ist, wenn sich die Breiten der entstehenden Bins nicht zu stark unterscheiden. Im Falle des oben aufgeführten Features „Capital Gain“, das eine sehr linkssteile Verteilung hat, erstreckt sich bei einer Aufteilung in Dezile ein Bin über den gesamten Wertebereich, während die anderen eine Breite von 0 aufweisen.

Demnach ist Diskretisierbarkeit gegeben, wenn die Datenpunkte einer Feature-Verteilung in gleich-große Intervalle mit ähnlicher Anzahl der darin liegenden Datenpunkte, oder in bspw. Dezile mit ähnlicher Breite aufgeteilt werden können.

Die ausgewählten XAI-Methodenimplementierungen verwenden EqualFrequency Binning mit einer standardmäßigen Aufteilung in vier (*Alibi* Implementierungen von Anchors und CFProto) bzw. 20 Bins (*PyALE*). Da die Anzahl der Bins angepasst werden kann, wird die Diskretisierbarkeit nachfolgend durch einen Vergleich der Bin-Grenzen von Dezilen ermittelt.

Bei einer Aufteilung des Wertebereiches in Dezile weist jeder Bin optimalerweise eine Breite von 10% (0.1) auf. Daher kann man sagen, dass ein Feature umso schlechter diskretisierbar ist, je höher die durchschnittliche Abweichung der Bin-Breiten zu diesem optimalen Verhältnis ist.

Zur Berechnung dieser Abweichung zu 0.1 kann der MAD auf dem Verhältnis der Bin-Breiten angewandt werden. Die Diskretisierbarkeit wird folglich durch die Subtraktion des MADs von der optimalen Bin-Breite ermittelt. Sollte der MAD größer als 0.1 sein, ist der resultierende Wert negativ, was auf die Existenz sehr großer Breitenunterschiede hindeutet. In diesem Fall gibt es mindestens einen sehr

großen Bin, der die Breite der anderen stark reduziert. Je höher dieser Wert ist, desto besser ist die Diskretisierbarkeit des Features.

Die ermittelten Diskretisierbarkeitswerte pro Feature werden anschließend, wie auch bei der Korrelationsberechnung, durch eine Kombination des Wertedurchschnitts mit dem MAD zu einem einzigen Eingabewert aggregiert. Wie bei der Korrelationsermittlung wird der MAD dabei halbiert, um Werte ähnlich menschlicher Einschätzungen zu erhalten. Beispiele der Berechnung der Diskretisierbarkeit sind in Anhang A.6.2 aufgeführt.

#### **4.3.3. Identifikation der Eingaben des Fuzzy-Systems**

Aktuell umfassen die Systemeingabeparameter sowohl die Ausschlusskriterien als auch die Kriterien, nach denen die Eignung der XAI-Methoden bewertet wird. Die Ausschlusskriterien beeinflussen die Methodeneignungen allerdings nicht, sondern entfernen die XAI-Methoden ggf. aufgrund ihrer Nicht-Ausführbarkeit aus der Empfehlungsergebnismenge. Daher werden diese Parameter nicht in das Fuzzy-Expertensystem des XAIRs gegeben, welches die Eignung jeder XAI-Methode durch die Max-Min-Inferenz ermittelt.

Um den Ausschluss einer nicht anwendbaren XAI-Methode zu gewährleisten, werden im Anschluss an das Fuzzy-System alle XAI-Methoden, die durch Nicht-Erfüllung der Voraussetzungen nicht anwendbar sind, aus der Ergebnismenge eliminiert. Dem Nutzer werden somit nur die geeignetsten, tatsächlich anwendbaren Methoden präsentiert.

Die Eingaben des Fuzzy-Expertensystems beschränken sich daher auf:

- Korrelation
- Korrelation der FOI
- Diskretisierbarkeit
- Diskretisierbarkeit der FOI
- Präferenzen der Leistung
- Anzahl der Features
- Zugriffszeit des Modells/der Vorhersagefunktion
- Vorbereitungsaufwand
- Präferenz einer globalen Erklärung
- Präferenz einer lokalen Erklärung
- Vorhandensein ordinaler Features

#### **4.3.4. Festlegung der Fuzzifizierung und Defuzzifizierung**

Nach der Identifikation der Eingabe- und Ausgabevariablen des Fuzzy-Expertensystems ist, wie in Kapitel 2.4.2 erwähnt, eine Festlegung und Aufteilung ihrer Fuzzy-Wertebereiche und eine Modellierung der Zugehörigkeitsfunktionen notwendig. In diesem Zuge wird eine Defuzzifizierungsstrategie festgelegt, welche aus der Fuzzy-Ergebnismenge einen festen, interpretierbaren Wert zurückgibt.

### Modellierung der Fuzzifizierungsfunktionen

Die Modellierung der vagen Zugehörigkeitsfunktionen und die Beurteilung der Mitgliedsgradwerte erfolgt aufgrund von Erfahrungen, persönlichen Einschätzungen und sprachlichen Gewohnheiten nach sachinhaltlichen Gegebenheiten (Böhme 1993, S. 5). Sie ist daher kontextabhängig und kann intersubjektiv variieren.

Für als Einschätzung empfangene Eingaben, z.B. für die Präferenzen der Leistung, wird ein Diskursuniversum  $X = [0; 10]$  definiert. Dieses wird in drei trianguläre Zugehörigkeitsmengen unterteilt, da eine feingranularere Aufteilung für diese grobe Einschätzung nicht notwendig ist.

Auch für die zweiwertigen Parameter (Existenz ordinaler Features und lokaler/globaler Erklärungsumfang) werden trianguläre Zugehörigkeitsfunktionen definiert, da zweiwertige bzw. Singleton Eingaben bei der verwendeten Implementierung, siehe Kapitel 5.1, nicht möglich sind.

Für die Variablen der Korrelation ist jeweils ein Diskursuniversum  $X = [0; 10]$  vorgesehen. Der in der Datenanalyse ermittelte Korrelationswert, der einen Maximalwert von 1 erreichen kann, wird dabei auf 10 skaliert. Diese Skalierung ist notwendig, da die verwendete Implementierung der Fuzzy-Logik Probleme beim Setzen der Grenzen der Zugehörigkeitsfunktionen auf Dezimalzahlen hat: Bei ihrer Erstellung wird jeder ganzzahligen Abszisse (jedem Wert des Diskursuniversums) anhand der gegebenen Funktionsform ein Ordinaten (ein Zugehörigkeitswert) zugewiesen. Das Setzen nicht-ganzzahliger Werte für die Funktionsform führt zu verfälschten Zugehörigkeitsfunktionen und resultierenden -werten. Dies wird am nachfolgenden Beispiel dreier trapezförmiger Zugehörigkeitsfunktionen, siehe Tabelle 4.4 und Abbildung 4.7, verdeutlicht.

Tabelle 4.4.: Beispielhafte problematische Definition der Zugehörigkeitsfunktionen mit Dezimalzahlen bei *scikit-fuzzy*

Fuzzy-Menge	Zugehörigkeitsfunktion	Funktionsverlauf (*)
<b>L</b>	(0, 0, 1, 2.5)	[1, 1, 0.333, 0, 0]
<b>M</b>	(1, 2, 3, 4)	[0, 0, 1, 1, 0]
<b>H</b>	(2.5, 4, 5, 5)	[0, 0, 0, 0.333, 1]

(\*) Vektoren der Zugehörigkeitswerte (y-Achse) je ganzzahligem Wert des Diskursuniversums (x-Achse)

Beim obigen Beispiel sollte eine crisp Eingabe von 2.5 die folgenden Zugehörigkeitswerte aufweisen:

$$\mu_L(2.5) = 0.0$$

$$\mu_M(2.5) = 1.0$$

$$\mu_H(2.5) = 0.0$$

Stattdessen liefern die verfälschten Zugehörigkeitsfunktionen die folgenden Werte:

$$\mu_L(2.5) = 0.167$$

$$\mu_M(2.5) = 1.0$$

$$\mu_H(2.5) = 0.167$$

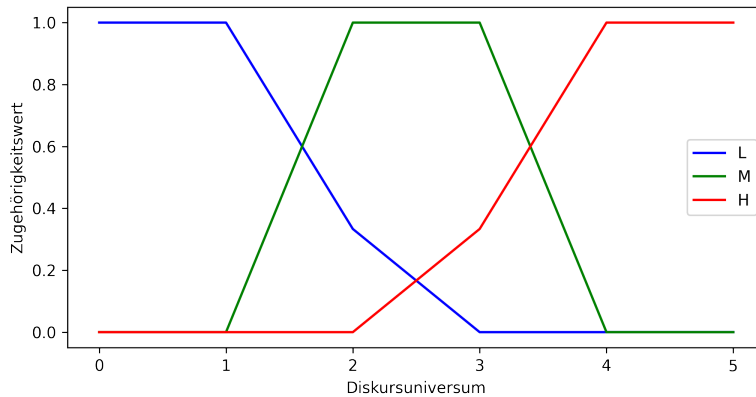


Abbildung 4.7.: Visualisierung der durch *scikit-fuzzy* verfälschten Zugehörigkeitsfunktionen der Tabelle 4.4

Durch die Skalierung auf 10 wird das geschilderte Problem behoben.

Die Bewertung der XAI-Methoden hinsichtlich der Korrelation findet für drei Ausprägungen statt, da die Literatur diesbezüglich keine genaueren Auskünfte gibt. Allerdings ist eine Einteilung der Eingabevariablen in fünf Zugehörigkeitsmengen vorgesehen, um spezifischere Hinweise bei sehr starken Korrelationen geben zu können. Für eine korrekte Aktivierung der Bewertungsregeln muss SCHWACH die Menge SEHR SCHWACH, und STARK die Menge SEHR STARK beinhalten.

Die Zugehörigkeitsfunktionen orientieren sich an den Korrelationsinterpretationen der Tabellen 4.2 (S. 46). Sie werden zu trapezförmigen Funktionen zusammengefasst, deren Kern aus den Überschneidungen der beiden Daumenregeln besteht, und ihre stützende Menge (vgl. Abbildung 2.2, Seite 14) bei dem minimalen und dem maximalen Korrelationswert endet.

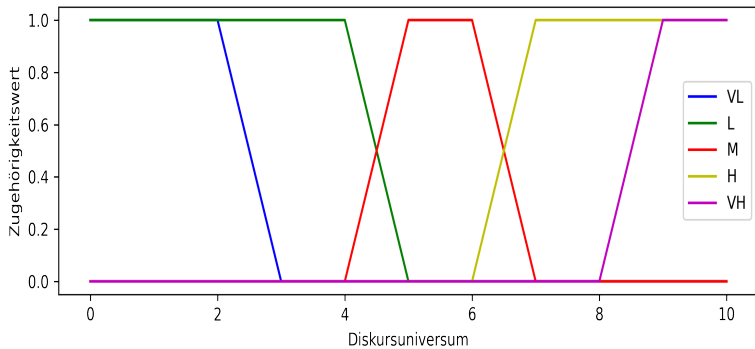
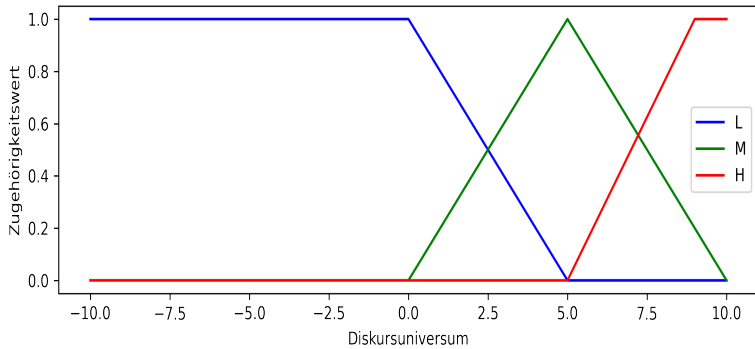
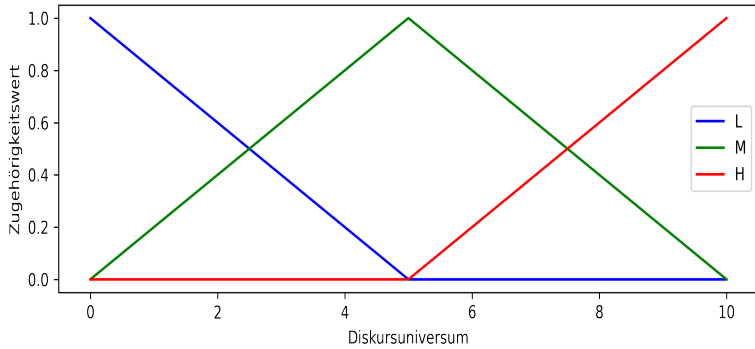
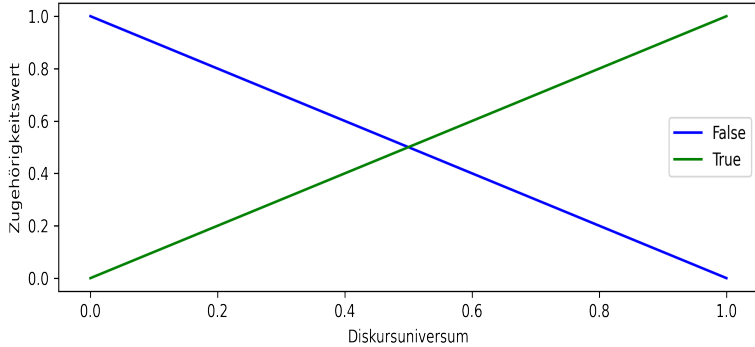
Die Unschärfe der Diskretisierbarkeit wird auf dem Diskursuniversum  $X = [-10; 10]$  eingeschätzt, da bei der in Kapitel 4.3.2 vorgestellten Berechnungsweise ihre Werte bei sehr schlechter Diskretisierbarkeit auch negativ sein können. Der Eingabeparameter wird hochskaliert, sodass er bei optimaler Diskretisierbarkeit den Wert 10 annimmt.

Es werden drei Ausprägungen mit triangulären Fuzzy-Mengen angenommen; eine feingranularere Unterteilung ist aufgrund der Nicht-Quantifizierbarkeit der Eingabeparameter und der Methodenempfehlung nicht möglich bzw. nicht notwendig.

Die Zugehörigkeitsfunktionen der Eingabevariablen des Fuzzy-Expertensystems sind in Tabelle 4.5 zusammenfassend aufgeführt.

Die Ausgaben des XAIRs spiegeln die Eignung der einzelnen XAI-Methoden wider. Daher wird für jede ausgewählte Methode eine Ausgabevariable mit einem einheitlichen Diskursuniversum  $X = [0; 10]$  erstellt. Der Wertebereich der Eignung wird in fünf Ausprägungen mit triangulären Zugehörigkeitsfunktionen unterteilt, siehe Abbildung 4.8: Ungeeignet (VL), eher ungeeignet (L), neutral (M), geeignet (H) und sehr gut geeignet (VH).

Tabelle 4.5.: Zugehörigkeitsfunktionen der Eingabevariablen des Fuzzy-Expertensystems

Kriterien	Eigenschaften der Zugehörigkeitsfunktionen	
Korrelation (der FOI)	 <p>Diskursuniversum: [0;10]</p> <p>Funktionsformen:  (0, 0, 2, 3)  (2, 3, 4, 5)  (4, 5, 6, 7)  (6, 7, 8, 9)  (8, 9, 10, 10)</p>	
Diskreti- sierbarkeit (der FOI)	 <p>Diskursuniversum: [-10;10]</p> <p>Funktionsformen:  (-10, -10, 0, 5)  (0, 5, 10)  (5, 9, 10, 10)</p>	
Leistungs- präferenz Anzahl der Features Dauer des Modell- zugriffs Vorbereitungs- aufwand	 <p>Diskursuniversum: [0;10]</p> <p>Funktionsformen:  (0, 0, 5)  (0, 5, 10)  (5, 10, 10)</p>	
Präferenz des Erklär- ungsumfangs (global/lokal) Vorhandensein ordinaler Features	 <p>Diskursuniversum: [0;10]</p> <p>Funktionsformen:  (0, 0, 1)  (0, 1, 1)</p>	

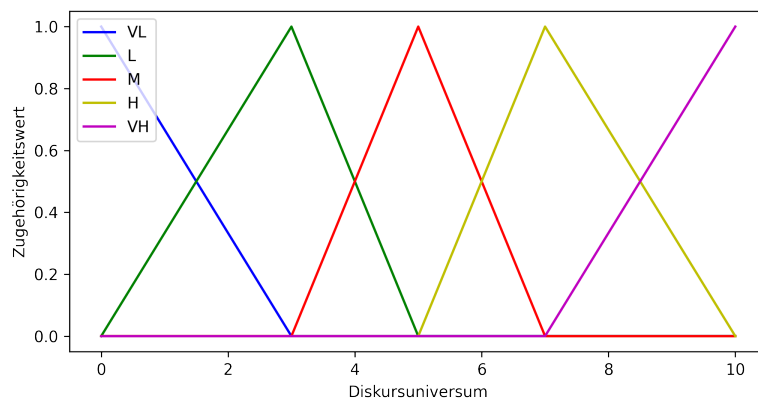


Abbildung 4.8.: Zugehörigkeitsfunktionen der Ausgabevariablen (Methodeneignungen) des Fuzzy-Expertensystems

### Wahl der Defuzzifizierungsstrategie

Aus der Fuzzy-Ergebnismenge muss ein crisp Ausgabewert ermittelt werden, der die Methodeneignung repräsentiert. Dafür wurden maxima- und verteilungsbasierte Defuzzifizierungsmethoden in Betracht gezogen.

Der häufigste Anwendungsfall von Maxima Methoden ist in wissensbasierten Fuzzy-Systemen (van Leekwijck & Kerre 1999). Bei der Ergebnisberechnung ignorieren diese allerdings alle nicht maximal aktivierten Regeln, womit das System seinen fuzzy Charakter verliert (Mamdani et al. 1984, S. 890). Dies ist beispielhaft in der generierten Eignungsbewertung von Anchors in Abbildung 4.9 zu sehen.

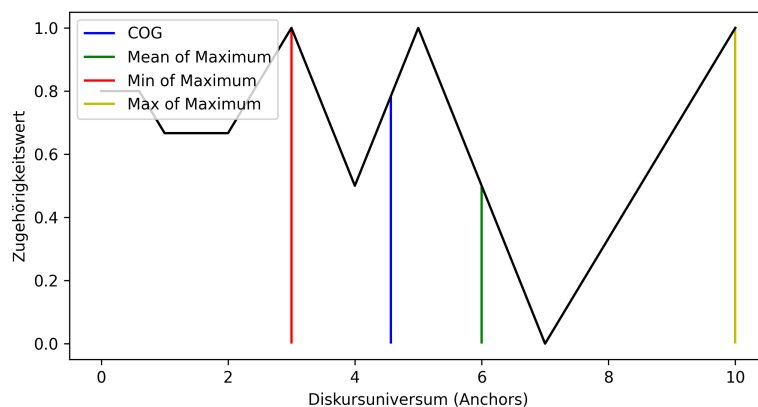


Abbildung 4.9.: Beispielhafte Ergebnisse betrachteter Defuzzifizierungsstrategien der Ausgabevariable „Anchors“

Die Verwendung einer solchen Defuzzifizierungsmethode innerhalb des XAIRs kann dazu führen, dass einzelne Kriterienbewertungen nicht in die Gesamtbewertung der XAI-Methode miteinfließen.

Die COG Methode weist zwar eine höhere Berechnungskomplexität auf, allerdings ist diese hauptsächlich im Kontext von Echtzeitregelungssystemen relevant und in diesem Fall aufgrund der simplen

Formen der Zugehörigkeitsfunktionen und einer überschaubaren Menge an Regeln und Ein- und Ausgaben zu vernachlässigen. Ein weiterer Nachteil der COG Methode ist, dass die scharfe Ausgabe durch eine Mittelung nur selten zu den Extremen des Bewertungsbereichs tendiert (Mamdani et al. 1984), d.h. in diesem Fall zu „sehr gut geeignet“ oder „sehr schlecht geeignet“. Dies ist für den XAIR ebenfalls irrelevant, da das Ergebnis der Eignung einer XAI-Methode relativ zur Eignung der anderen gesehen wird.

Daher wird für die Defuzzifizierung des Empfehlungsergebnisses die COG Strategie verwendet, die alle Regeln für die Eignungsbewertung der XAI-Methode berücksichtigt.

#### 4.3.5. Erstellung der Regelbasis

Die Fuzzy-Expertenregeln der Wissensbasis drücken Fachwissen über die Zusammenhänge zwischen den identifizierten Kriterien und der Eignung der XAI-Methoden aus.

Für die Erstellung der Fuzzy-Wissensbasis werden zunächst die Eignungen der XAI-Methoden hinsichtlich der einzelnen Voraussetzungen und Kriterien ermittelt. Da diese teilweise auch von Kombinationen bestimmter Kriterien beeinflusst werden, werden anschließend weitere Regeln definiert.

#### Bewertung der XAI-Methoden

Die Voraussetzungen der XAI-Methoden werden, mit den in Anhang A.3 aufgeführten Formulierungen, binär in Tabelle 4.6 bewertet.

Tabelle 4.6.: Bewertung der XAI-Methoden hinsichtlich der Voraussetzungen

Kriteriumsname	Methode					
	PDP + ICE	ALE	PFI	SHAP	Anchors	CFProto
Verfügbarkeit des Modells	0	1	1	0	0	0
Klassifikationsaufgabe	0	0	0	0	1	1
Erhalt der Klassenwahrscheinlichkeiten	0	0	0	0	0	1
Zugriff auf Labels	0	0	1	0	0	0
Zugriff auf Preprocessing Operationen	0	0	0	0	1	1

Sollte das jeweilige Kriterium für die Anwendung notwendig sein, wird die Methode mit „1“ (ja) bewertet. Wie bereits erwähnt werden anhand dieser Bewertungen die XAI-Methoden vor der Empfehlungspräsentation aus der Ergebnismenge entfernt. Es gilt: Wenn die Parametereingabe des Nutzers größer gleich der Methodenbewertung ist, ist eine Anwendung möglich. Sollte sie kleiner als die Methodenbewertung sein, ist die Voraussetzung für die Anwendung nicht gegeben und sie ist nicht anwendbar.

Die fuzzy Regelbasis bzgl. der Methodenbewertung ist in Tabelle 4.7 dargestellt, in der die Methoden-eignung abhängig von der Größe/Stärke der mehrwertigen Eingabekriterien bewertet ist. Die Prämisse einer Regel ist die linguistische Variable und ihr Term (Spalte 1 und 2), und die Konklusion die jeweilig aufgeführten Methodenbewertungen. Bewertungskonzepte und Hinweise zur Interpretation der

Tabelle sind nachfolgend aufgeführt.<sup>1</sup>

Tabelle 4.7.: Bewertung der XAI-Methoden bzgl. der fuzzy Kriterien

Kriterium	Eingabe	Methode					
		PDP + ICE	ALE	PFI	SHAP	Anchors	CFProto
Korrelation	L	VH	L	-	-	-	-
	M	-	-	L	L	L (60%)	L (60%)
	H	L	VH	VL	VL	L	L
Korrelation der FOI	L	VH	L	-	-	-	-
	M	L	H	-	-	-	-
	H	VL	VH	L	L	L (60%)	L (60%)
Diskretisierbarkeit	L	-	L (60%)	-	-	VL	-
	M	-	-	-	-	L	-
	H	-	-	-	-	-	-
Diskretisierbarkeit der FOI	L	-	L	-	-	VL	-
	M	-	-	-	-	L	-
	H	-	-	-	-	-	-
Anzahl der Features	L	-	-	-	-	-	-
	M	-	-	-	-	L	-
	H	-	-	L	VL	VL	L
Zugriffszeit Modell/ Vorhersagefunktion	L	-	-	-	-	-	-
	M	-	-	-	-	L	VL
	H	-	-	L	L	VL	VL
Aufwand Vorbereitung	L	-	-	L	L	VL	VL
	M	-	-	-	-	L	L
	H	-	-	-	-	-	-
Globale Erklärung	0	-	-	-	-	-	-
	1	VH	VH	VH	VH	-	-
Lokale Erklärung	0	-	-	-	-	-	-
	1	VH	-	-	VH	VH	VH

### Hinweise zu den Bewertung der XAI-Methoden

Die Nutzerpräferenzen bzgl. der Leistung und das Vorhandensein ordinaler Features im Datensatz spielen in Kombinationen mit anderen Kriterien eine Rolle. Da sie eigenständig keinen Einfluss auf die Methodeneignung haben, sind sie nicht in der Tabelle 4.7 aufgeführt.

Für die Fuzzy-Eingabeparameter erfolgt die Bewertung in den vier linguistischen Termen „VL“, „L“, „H“ und „VH“, wobei bewusst auf eine Bewertung mit „M“ verzichtet wurde. Das hat zwei Gründe:

Abhängig von der Ausprägung der Kriterien ist für eine graduelle Abstufung der Eignung eine solche Bewertung zwar manchmal gewünscht, jedoch sollte sie in Tabelle 4.7 in jedem Fall als güterreduzierend gesehen werden. Falls eine XAI-Methode aufgrund der anderen Kriterien schlechter als „M“ bewertet wäre, würde sich eine solche Wertung positiv auf die Eignungseinschätzung auswirken. Daher

<sup>1</sup>An dieser Stelle wird erneut auf die Unschärfe der Literatur und die Subjektivität der Einschätzungen hingewiesen, da die Methodenbeurteilung aktuell nur durch die Autorin nach einer Literaturrecherche und eigenen Implementierungsversuchen durchgeführt wurde. Besonders die Einschätzung des Vorbereitungsaufwands ist sehr subjektiv und wesentlich von den datenwissenschaftlichen Kenntnissen des Nutzers abhängig.



wird statt einer „M“- eine „L“-Bewertung mit reduziertem Regelgewicht gegeben. Eine mittelstarke Korrelation führt z.B. zu einer „etwas schlechteren“ Bewertung von Anchors (siehe Tabelle 4.7), da die Regel die Konklusion nur mit einer Stärke von 60% bewertet. Folglich mindert diese Regel ihre Eignung von Anchors, wobei sie eine bereits schlechte Bewertung bei  $\mu_L(\text{Anchors}) > 0.6$  belässt, und sie bei  $\mu_L(\text{Anchors}) < 0.6$  verstärkt.

Ein weiterer Faktor, der ebenfalls im Kontext der Defuzzifizierung betrachtet werden muss, ist das Setzen einer initialen Bewertung der XAI-Methode. Das ist obligatorisch, da das Fuzzy-System verlangt, dass bei der Ergebnisausgabe für alle Ausgabeparameter ein definierter Zustand vorhanden ist. Daher findet eine initiale Bewertung aller XAI-Methoden durch Aktivierung einer Regel statt.

Dafür wurde eine „VH“-Bewertung in Betracht gezogen, da die Kriterien fast ausschließlich für Eignungsreduzierungen verantwortlich sind. Ohne ihr Vorhandensein eignen sich alle Methoden „sehr gut“. Dies führt allerdings im Zusammenhang mit der COG Defuzzifizierung zu Problemen: Sollte für eine XAI-Methode nur eine Regel aktiviert werden, die sich bspw. mit „VL“ sehr negativ auf ihre Bewertung auswirkt, pendelt sich die crisp Ausgabe in der Mitte ein. Eine Korrektur des Ergebniswerts nach oben durch eine weitere „VH“ Bewertung ist dann nicht mehr möglich.

Dass negative Bewertungen nicht zu sehr ins Gewicht fallen und um Steigerungspotential nach oben zu bieten, wird daher ein initialer Wert von „M“ gewählt. Weitere Bewertungen der Methode mit „M“ innerhalb der Wissensbasis werden dadurch, aufgrund der bereits vollen Mitgliedschaft zu dieser Zugehörigkeitsmenge, nicht mehr berücksichtigt.

Ein resultierender, crisp Empfehlungswert von 5 („M“) sagt folglich nicht zwangsläufig eine mittelmäßige, sondern eher eine neutrale Eignung aus und sollte daher nicht als schlecht interpretiert werden. Er kann durch einen Ausgleich positiver und negativer Bewertungen oder durch die Abwesenheit von eignungsreduzierenden Kriterienbewertungen entstehen.

Kriterienwerte, die auf die Methodeneignung keinen Einfluss haben, sollen außerdem weder positiv noch negativ ins Gewicht fallen, da sich bspw. Anchors bei schwacher Korrelation des Datensatzes nicht besser eignet. Diese irrelevanten Auswirkungen sind in der Tabelle 4.7 mit „-“ gekennzeichnet. Wie außerdem ersichtlich wirkt sich eine Präferenz bzgl. des Erklärungsformates ausschließlich positiv auf die Methodengüte aus.

### Hinweise zur Gewichtung der Regeln

Es wird außerdem davon ausgegangen, dass die Visualisierungsmethoden zur Vermeidung von Diskriminierung primär auf die FOI angewendet werden. Da die Visualisierung anderer Features ebenfalls möglich/gewünscht ist, wird die Methode auch für die Parameter, die sich auf den Gesamtdatensatz beziehen, bewertet. Diese sind für die Methodenanwendbarkeit allerdings aus oben genanntem Grund weniger wichtig als die FOI-spezifischen, weshalb sie hinsichtlich der Visualisierungsmethoden geringer gewichtet werden. Da die Gesamtparameter zudem den FOI-spezifischen Wert beinhalten, werden analog Nicht-Visualisierungsmethoden für die Ausprägung der FOI-spezifischen geringer gewichtet. Diese Regeln mit einer geringeren Gewichtung von 70% der Konklusion werden in Tabelle 4.7 mit grau hinterlegten Zellen dargestellt.

**Definition kombinierter Regeln**

Bei Kriterien, die keine eigenständige Auswirkung auf die Eignung einer XAI-Methode haben, kann unterschieden werden zwischen:

- Kriterien, die durch andere erst relevant („aktiviert“) werden
- Kriterien, die abhängig von den Eingabewerten anderer die Methodeneignung in unterschiedlichem Ausmaß beeinflussen

Die, die als Vorbedingung für die Aktivierung anderer dienen, werden in den Fuzzy-Regeln mit dem Fuzzy-Operator AND verknüpft. Falls keine FOI im Datensatz vorhanden sein sollten, werden die FOI-spezifischen Regeln so nicht aktiviert. Außerdem werden die Auswirkungen der Kriterien, welche die Methodenleistung beeinflussen, abhängig von der gegebenen Leistungspräferenz gewichtet.

Für Eigenschaften, die miteinander interagieren und neben den normalen Bewertungen aus der oben aufgeführten Tabelle 4.7 zusätzlich Interaktionseffekte für bestimmte Methoden haben, werden zusätzlich spezifische Regeln erstellt. Diese betreffen die Bewertung der XAI-Methoden ALE und CFProto.

Auf die Eignung von ALE hat bspw. eine mittelmäßige oder gute Diskretisierbarkeit (der FOI) nur in Kombination mit der Korrelation eine negative Auswirkung. Wenn beide Werte „H“ entsprechen, wird die Güte von ALE abhängig von der Höhe der Diskretisierbarkeit durch die AND-Verknüpfung nochmals explizit positiver bewertet.

Falls bspw.  $\mu_H(\text{corr}) = 0.4$  und  $\mu_H(\text{discr}) = 0.8$ , führt das zu einer Bewertung von:  $\mu_{VH}(ALE) = 0.7$ .

Abhängig von der Existenz ordinaler Features wird die Güte von CFProto bei schlechter Diskretisierbarkeit bzw. niedriger Korrelation reduziert. Die Distanzmessung, die für die Ermittlung der natürlichen Ordnung erforderlich ist, diskretisiert nämlich numerische Features. Außerdem geht sie davon aus, dass Korrelationen zwischen den ordinalen und den anderen Features existieren (Le & Ho 2005).

Das gesamte Fuzzy-Regelwerk ist in Anhang A.7 aufgeführt.

**4.3.6. Aufbau der grafischen Benutzeroberfläche (GUI)**

In diesem Kapitel wird auf die Konzeption der GUI der Web-Anwendung eingegangen. Dafür wurde gemäß DSR regelmäßig Feedback bzgl. der GUI in Iterationen eingeholt, wobei Personen befragt wurden, die der identifizierten Zielgruppe des Systems entsprechen.

Für die Erstellung wurden die in Kapitel 3.3 erhobenen Anforderungen berücksichtigt: Die GUI soll leicht verständlich und somit benutzerfreundlich (NFR3) sein. Dem Nutzer, der keine tiefen ML-Kenntnisse aufweist, sollen die Empfehlungsergebnisse übersichtlich präsentiert (FR1) und nachvollziehbar begründet (FR2) werden. Der XAIR soll näher auf die am besten geeignete XAI-Methode eingehen (FR5) und durch Präsentieren der für eine Implementierung wichtigen Informationen die Bereitschaft der Ausführung der Methode erhöhen (FR8.1). Außerdem wird im Sinne von FR8 bei der

Auswahl und der Strukturierung der darzustellenden Informationen darauf geachtet, dass der XAIR einen Lerneffekt für den Nutzer hat: Es soll als Nachschlagewerk für gepflegte Methoden und Implementierungen dienen, generelle Informationen zur Förderung der Nachvollziehbarkeit bereitstellen (FR8.2) und dem Nutzer die Möglichkeit geben, herauszufinden, wie sich Änderungen der Eingaben auf den Methodenvorschlag auswirken (FR7).

Eine Schwierigkeit bei der Konzeption der GUI war die Erfüllung von FR1, der übersichtlichen Ergebnispräsentation.

Da keine konkrete Angabe der Eignung in Prozent gegeben werden kann, wurde zunächst ein statisches „Siegerpodest“ für die Aufführung der Empfehlungen in Betracht gezogen. Allerdings sind so die Differenzen der Eignungsbewertungen nicht ersichtlich und diese somit nicht vergleichbar. Die Differenz ist bei einem Balken- bzw. Säulendiagramm ebenfalls nicht offensichtlich erkennbar, da die Unterschiede der resultierenden crisp Eignungswertungen sehr klein sind. Die iterative Evaluation wies darauf hin, dass eine Beschriftung der Balken mit ihren tatsächlichen crisp Werten dem Nutzer ohne Kenntnisse der Funktionsweise eines Fuzzy-Expertensystems keinen Mehrwert bietet und ihn eher irritiert.

Da die Eignung der XAI-Methoden relativ zu denen der anderen gesehen wird, wurde ein kreisförmiges, proportionales Flächendiagramm erwogen.

Diese Diagrammart eignet sich sehr gut für den Vergleich von Werten und die Darstellung von Proportionen, um einen schnellen Gesamtüberblick über die relativen Größen der Daten zu erhalten. Da die Werte der Daten schwer zu schätzen sind, hat es ausschließlich Kommunikationszwecke und keine analytischen. (The Data Visualisation Catalogue 2019)

Obwohl diverse Transformationsversuche des crisp Eignungswertes  $x$  bei der Berechnung der Flächengröße getestet wurden (z.B.  $x^2$ ,  $\pi^x$ ), sind allerdings auch dafür die Differenzen der Flächengrößen zu gering, um einen signifikanten Eignungsunterschied zu erkennen.

Letztendlich wurde eine relative Eignungsdarstellung in Form eines Säulendiagramms gewählt. Die Höhe der Säulen wird dabei abhängig vom höchsten Eignungswert, der die Höhe 100% erhält, und dem niedrigsten (Säulenhöhe 1%) skaliert. Das liefert gut unterscheidbare Säulenhöhen und erlaubt den relativen Eignungsvergleich der einzelnen XAI-Methoden.

Abhängig von der Zielgruppe oder von der Häufigkeit der Anwendung des XAIRs sind nicht alle verfügbaren Informationen für den Nutzer interessant (vgl. FR5). Daher wurde im Laufe der Iterationen die ursprünglich geplante Single-Page Anwendung zu einer Multi-Page Anwendung umstrukturiert und in überspringbare Teile segmentiert.

Die übersichtliche, thematische Unterteilung in einzelne Seiten findet anhand der nachfolgend aufgeführten Fragen statt. Die eingerückten Fragen strukturieren dabei den jeweiligen Seiteninhalt.

- **Eingabeseite:** Welche Eingaben werden für eine Empfehlung benötigt?
- **Ergebnisseite:** Welche XAI-Methoden werden empfohlen?
- **Erklärungsseite:** Warum wurden diese XAI-Methoden empfohlen?
  - Anhand welcher Eingaben wurden die Eignungen der XAI-Methoden ermittelt?
  - Welche XAI-Methoden können aufgrund der Eingaben nicht angewendet werden?

- Wie wurden die Eignungen der anwendbaren XAI-Methoden anhand der einzelnen Eingaben bewertet?
- Welche XAI-Methoden wurden für eine Empfehlung in Betracht gezogen?
- **Detailseite der empfohlenen XAI-Methode M:** Welche XAI-Methode wird empfohlen?
  - Wie wird M taxonomisch eingeordnet?
  - Welche Frage beantwortet M? (Anhand eines Beispiels)
  - Wie funktioniert M?
  - Was ist das Ergebnis von M?
  - Welche Frage beantwortet M *nicht*?
- **Detailseite der empfohlenen Implementierung:** Wie kann M angewendet werden?
  - Welche Implementierung wird empfohlen?
  - Wo ist die Dokumentation und der Sourcecode zu finden?
  - Was benötigt man für die Anwendung?
  - Was ist das Ergebnis der Implementierung?
  - Was muss man beachten?
    - \* Hinweise zur Anwendung der XAI-Methode
    - \* Hinweise zur Parameterauswahl für die Implementierung
- **Allgemeine Empfehlungsseite:** Wie soll man weiter vorgehen?

Die Seiten können über eine Navigationsleiste gemäß der dargestellten, empfohlenen Reihenfolge Schritt für Schritt durchgegangen, oder gezielt angeklickt bzw. übersprungen werden. Nur die Detailseite der höchstbewerteten, empfohlenen XAI-Methode wird in dieser Leiste angezeigt; die Detailseiten der anderen Methoden und Implementierungen sind dennoch verfügbar und aufrufbar.

Zusätzlich gibt es eine Informationsseite („FAQ“), welche von jeder Seite aus erreichbar ist und Antworten auf die folgenden, allgemeinen Fragen beinhaltet:

- Warum sollte man XAI verwenden?
- Warum sollte man den XAIR verwenden?
- Was bedeutet es, wenn eine XAI-Methode geeignet ist?
- Wie funktioniert der XAIR?
- Welche XAI-Methoden werden für eine Empfehlung in Betracht gezogen?
- Welches Datenformat soll für die Parametereingabe berücksichtigt werden?
- Was bedeutet „Diskretisierbarkeit“?

Die Screenshots der Seiten sind in Anhang A.8 aufgeführt.

## 5. Implementierung des XAIRs

In diesem Kapitel werden zunächst die Technologien aufgeführt, die für die Umsetzung des prototypischen XAIR verwendet wurden. Da die für die Implementierung verwendeten Konzepte bereits ausführlich im vorigen Kapitel eruiert wurden, wird im Anschluss daran nur auf ausgewählte Implementierungsdetails eingegangen.

### 5.1. Verwendete Technologien

Das Backend des XAI-Empfehlungssystems wird in Python (Python Software Foundation 2020) umgesetzt, was eine leichte Integrierbarkeit in eine Python-basierte Kubeflow ML-Pipeline nach Anforderung NFR1 gewährleistet.

Zur Umsetzung der Web-Anwendung wurde das populäre Framework *Django* (Django Software Foundation 2021) in Betracht gezogen, welches unter anderem Unterstützung für bspw. das Rendern von HTML Seiten oder für eine vereinfachte Datenbankbindung bietet. Allerdings ist eine Trennung der GUI von der Anwendungslogik gewünscht, um die Integration in eine ML-Pipeline, ihre unabhängige Entwicklung und eine bessere Wartbarkeit (NFR4) zu ermöglichen.

Daher wird die backendseitige Schnittstelle zum Erhalt einer Methodenempfehlungen mit dem leichtgewichtigen Web-Anwendungs-Framework *Flask* (Ronacher 2020) erstellt. Zur Implementierung des Fuzzy-Expertensystems wird die auf GitHub öffentlich bereitgestellte Fuzzy-Logik Toolbox *scikit-fuzzy* (Warner 2019) aufgrund ihrer Popularität gewählt.

Für das Frontend wird eine *React* (Facebook Inc. 2020) Anwendung mit dem Framework *GatsbyJS* (Gatsby Inc. 2021) in TypeScript implementiert. Durch das Vorladen benötigter Ressourcen der Anwendungsseiten und ein serverseitiges Rendern statischer Seiten ist es sehr performant. Dieses serverseitige Vorrendern eignet sich zudem für die Bereitstellung der Methoden- und Implementierungsdetailseiten. Ihre Inhalte sind in methodenspezifischen JSON Dateien hinterlegt und werden beim Bau der Anwendung unter Angabe eines Templates erstellt.

### 5.2. Ausgewählte Implementierungsdetails

Nachfolgend werden einige ausgewählte Implementierungsdetails vorgestellt. Diese umfassen bestimmte, in Kapitel 4 noch nicht genannte Aspekte der Konfiguration, die zur Umsetzung des Konzepts notwendig sind. Außerdem werden nähere Einblicke in die Umsetzung der aus den Anforderungen abgeleiteten Nutzerwünsche gegeben.

#### 5.2.1. Konfiguration

Für einen minimalen Wartungsaufwand wurden wenige, einfach editier- und ergänzbare Konfigurationsdateien erstellt. Eine generelle System-Konfigurationsdatei (*config.json*) verwaltet die Dateipfade

benötigter Ressourcen und die Parameter des Fuzzy-Expertensystems. Neben den reduzierten Regelgewichten (siehe Kapitel 4.3.5) werden dort die Diskursuniversen und die Zugehörigkeitsfunktionen definiert. Der Gebrauch benutzerdefinierter Werte für bspw. neue Eingabekriterien ist dennoch möglich.

Die Ein- und Ausgabevariablen des Fuzzy-Expertensystems werden in jeweils eigenen Dateien, *antecedent\_config.json* und *consequent\_config.json*, spezifiziert. Für jede Eingabevariable sind unter anderem die Parameterart (Ausschluss- oder Bewertungskriterium) und der Datentyp definiert. Außerdem enthält *antecedent\_config.json* die für das Frontend benötigten Angaben, wie bspw. Name- und Bewertungslabels und ihren voreingestellten, sowie den minimalen und maximalen Wert.

Die Fuzzy-Regelbasis wird anhand von drei Dateien erstellt. In einer JSON Datei (*custom\_rules.json*) werden Regeln definiert, deren Prämisse aus einer Kombination von Kriterien besteht. Die anderen, welche den Ausschlusskriterien (Tabelle 4.6, S. 54) und den eignungsbeeinflussenden Kriterien (Tabelle 4.7, S. 55) entsprechen, werden aus zwei CSV-Dateien extrahiert und so in die Regelbasis mitaufgenommen. Obwohl innerhalb der zuerst genannten JSON Datei alle Regeln, die keinen Methodenausschluss bewirken, zentral aufgeführt werden können, wurde das CSV Dateiformat gewählt. Dieses ist mithilfe von Python leicht editier- und visualisierbar und somit übersichtlicher und nutzerfreundlicher.

Beim Bau der Anwendung findet die Initialisierung des XAIRs im Sinne der Trennung der Verantwortlichkeiten (Separation of Concerns) durch eine eigene Komponente, den *XAIRInitializer*, statt. Er bereitet alle vom XAIR benötigten Ein- und Ausgabevariablen und die Regeln vor, und prüft die Konsistenz der Methodenbewertungen in den CSV-Dateien mit den definierten XAI-Methoden und Kriterien.

Die schematische Komponentenarchitektur des XAIRs ist in Abbildung 5.1 dargestellt. Die Aufgaben des *XAIRInitializers* bzgl. des zur Verfügungstellens der Konfiguration sind durch die gestrichelten Linien angedeutet.

Die initiale Bewertung aller XAI-Methoden mit „M“ findet, wie bereits erwähnt, in Form der Aktivierung einer Regel statt, wofür eine weitere Eingabevariable (*init*) angelegt wird. Da die *scikit-fuzzy* Implementierung keine zu spärliche Regelbasis für das Fuzzy-System erlaubt, muss für jede Ausprägung (z.B. „L“, „M“, „H“) jedes Kriteriums mindestens eine Regel vorliegen. Daher wird neben den ausgewählten XAI-Methoden eine Nulloperations-Ausgabevariable definiert (*NOOP*). Diese wird bei einer Eingabe bewertet, die keinerlei Eignungsreduzierung verursacht.

Falls der Nutzer keine FOI angibt, werden FOI-spezifische Regeln nicht aktiviert. Dafür wird systemintern eine weitere Eingabevariable *foi\_available* definiert.

### 5.2.2. Erweiterbarkeit hinsichtlich neuer Kriterien und XAI-Methoden

Um eine einfache Erweiterung oder Veränderung der Wissensbasis zu gewährleisten, wird ein Jupyter Notebook für das Hinzufügen neuer (Ausschluss-)Kriterien und XAI-Methoden erstellt (*knowledge\_acquisition.ipynb*, siehe Abbildung 5.1). Jupyter Notebooks sind interaktive Testbereiche, die Code,

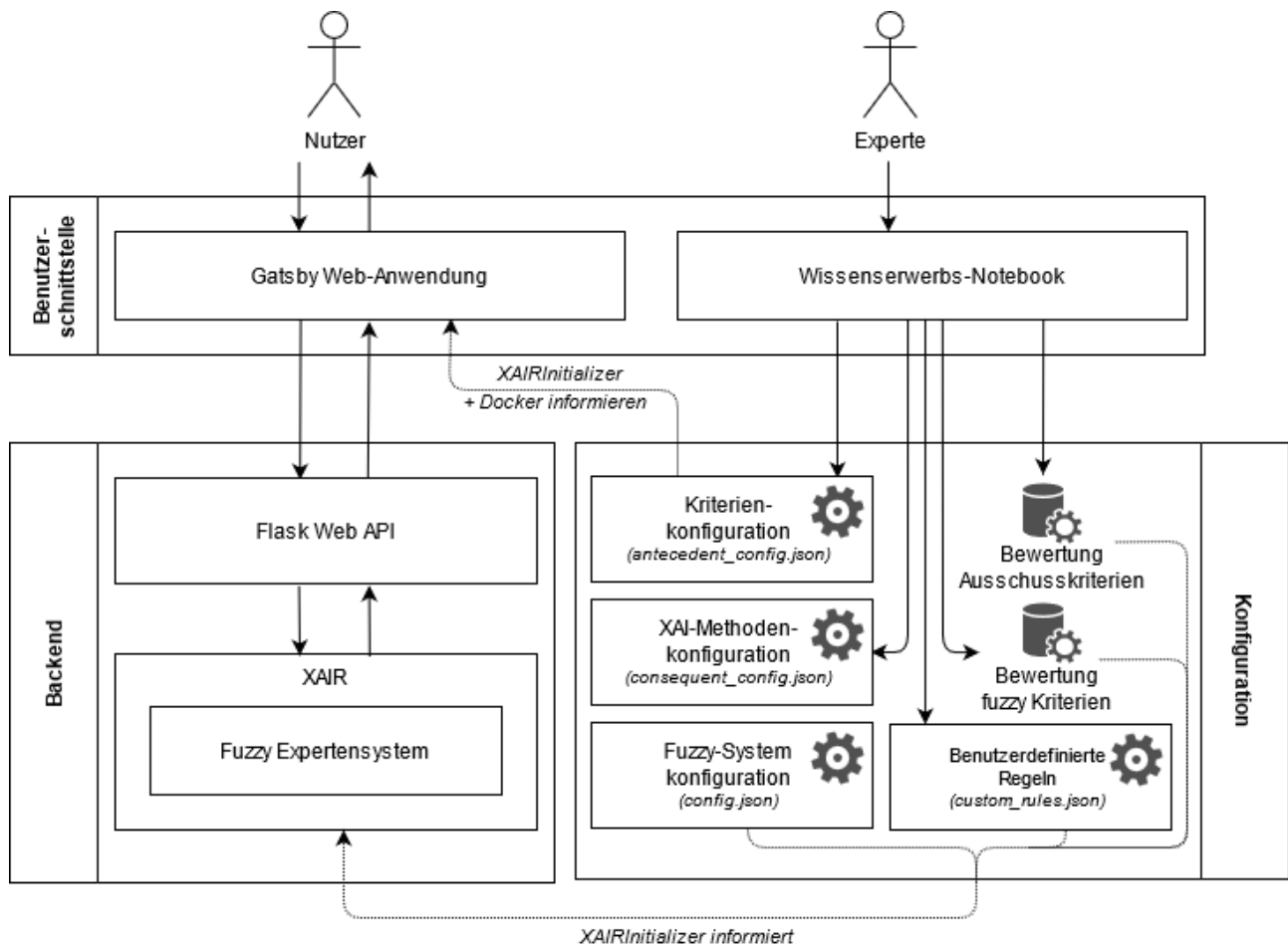


Abbildung 5.1.: Schematische Komponentenarchitektur des XAIRs

Visualisierungen und Text enthalten können und browser-basierte, interaktive Shells für verschiedene Programmiersprachen wie bspw. für Python bieten (La Vigne 2018).

Die benötigten Methoden- bzw. Kriterieneingaben werden über interaktive GUI Eingabefelder entgegengenommen. Der Aufbau des Notebooks ist dabei funktional und es wurde keinen Wert auf eine ansprechende Aufbereitung gelegt. Der Nutzer wird durch den gewünschten Hinzufüge- bzw. Editier-Prozess hindurchgeführt, weshalb lediglich eine sequenzielle Ausführung aller Code-Felder notwendig ist. Der Experte sollte folglich für die Formalisierung des Wissens zumindest in der Lage sein, das Jupyter Notebook auszuführen, oder einen damit vertrauten Wissensingenieur zu Rate ziehen.

Für das Hinzufügen einer XAI-Methode ist außerdem die Bereitstellung von detaillierteren Informationen innerhalb der Web-Anwendung vorgesehen. Dafür muss eine JSON Datei in der Frontend-Anwendung definiert werden, deren Vollständigkeit mithilfe eines zur Verfügung gestellten JSON Schemas validiert werden kann. Weitere Informationen hierzu sind ebenfalls im Jupyter Notebook zu finden.

Um die Konfigurationsänderungen zu übernehmen, muss beim nächsten Anwendungsstart ein reload-Flag gesetzt werden. Alternativ ist die Ausführung eines vorbereiteten Python Scripts möglich, welches den *XAIRInitializer* zum Nachladen der Ressourcen auffordert.

Beim Bau der Anwendung mit Docker wird die Konfigurationsdatei der Eingaben in das Frontend-

Projekt kopiert. Dadurch sind die produktiv eingesetzten Front- und Backend-Teile der Anwendung bzgl. der Kriterien stets konsistent und der Code der GUI muss nicht weiter angepasst werden.

### 5.2.3. Validierung der Eingabeparameter

Zur Erfüllung von FR3.2 wird eine front- und backendseitige Validierung der Eingabedaten umgesetzt. Dank der Echtzeitvalidierung der Eingabefelder mit der *React* Bibliothek *Formik* (Formium Inc. 2020) sind ungültige Anfragen an das Backend bei Benutzung der Web-Anwendung unmöglich.

Für eine automatisierte Verwendung des Empfehlungssystems muss sichergestellt werden, dass die Eingaben auch bei POST Anfragen anderen Ursprungs an die *Flask* API, oder bei direktem Aufruf des XAIRs valide sind. Dies geschieht mit einer leichtgewichtigen Datenvalidierungskomponente der Python Bibliothek *Cerberus* (Iarocci 2019). Ihr Validierungsregelschema wird entsprechend der konfigurierten Eingabe-Datentypen bei der Systeminitialisierung erstellt.

### 5.2.4. Nachvollziehbarkeit der Empfehlungsergebnisse

Erklärungen zu gegebenen Empfehlungen sind ein wichtiger Faktor zur Steigerung von Benutzerakzeptanz und Vertrauen in Empfehlungssysteme (Ziegler & Loepp 2019, S. 19). Die Ausgabe der für die Empfehlungsgenerierung aktivierten Regeln ist mit *scikit-fuzzy* allerdings nicht möglich. Um die Nachvollziehbarkeit zu fördern, wird daher die entsprechende, die Berechnungen durchführende Klasse des Fuzzy-Expertensystems um diese Funktionalität erweitert.

Sofern alle Terme der Prämisse erfüllt sind, wird die jeweilige Regel ausgegeben. Dabei wird bewusst auf die Berücksichtigung der Erfüllungsgrade der Prämisse, und somit auf den Erhalt der Gewichtung der Regel verzichtet. Die Angabe der Gewichtung erscheint dem Nutzer ohne Kenntnis der Max-Min-Inferenz willkürlich und ein nachvollziehbares Erklären für einen schnellen Empfehlungserhalt zu umfangreich und ggf. überfordernd.

In einer Tabelle in der Web-Anwendung werden die aktivierten Regeln nach Eingabekriterien sortiert aufgeführt und eine mögliche Mehrfachbelegung (durch Zugehörigkeit der Eingabe zu mehreren Fuzzy-Mengen) erläutert.



## 6. Evaluation

Um die Nutzbarkeit, Qualität und Wirksamkeit des XAIRs zu zeigen, wird der implementierte Prototyp nach dem Ansatz des Requirements Driven DSR anhand der Erfüllung der Anforderungen beurteilt. Mithilfe von semistrukturierten Interviews sollen die subjektiven Anforderungen und Systemziele des XAIR evaluiert, und konstruktives Feedback gesammelt werden.

Nachfolgend wird der Aufbau und die Durchführung der Interviews beschrieben und anschließend die Ergebnisse der Evaluation zusammengefasst und diskutiert.

### 6.1. Aufbau und Durchführung

Die qualitative Nutzerbefragung im Rahmen der Evaluation findet mit denselben Personen statt, die bereits an der Anforderungsanalyse teilgenommen haben. Diese sind daher mit dem Konzept und den Zielen des XAIRs vertraut. Obwohl ihr Feedback teilweise bereits für frühere Iterationsergebnisse eingeholt wurde, findet die hier aufgeführte, finale Evaluation mit einem, den Nutzern in dieser Form unbekannten Stand der Web-Anwendung statt.

Der Ablauf eines Interviews sieht wie folgt aus: Vor Beginn soll der Nutzer den XAIR für einen beliebigen, konkreten Anwendungsfall nach geeigneten XAI-Methoden befragen. Alle Interviewteilnehmer greifen dabei auf den bereitgestellten, konstruierten Titanic-Anwendungsfall zurück, der auf dem Titanic Datensatz (Kaggle Inc. 2021) basiert und in Anhang A.9 aufgeführt ist. Die Anwendung findet eigenständig und unabhängig statt und wird zeitlich nicht beschränkt. Dadurch können sich die Personen für eine Konsultation so viel Zeit nehmen, wie sie für dieses konstruierte, folglich nicht priorisierte Szenario investieren möchten. Allerdings wird eine Nutzung für mindestens 10 Minuten empfohlen, um einen ausreichenden Eindruck von dem XAIR zu gewinnen.

Anschließend folgt ein halbstrukturiertes Online-Einzelinterview, das sich an den Fragen des Leitfadens in Anhang A.10 orientiert. Die Fragen sind angelehnt an die von Miranda et al. (2011) identifizierten und nach beteiligten Entitäten (Nutzer, System, Organisation) organisierten Bewertungskriterien für Expertensysteme. Der Fokus des Interviews liegt dabei auf dem Nutzen (Einfachheit, erforderlicher Aufwand, Zufriedenheit) und der Ergebnisqualität (Vertrauen und Grad des Verständnisses der Entscheidungen) für den Nutzer. Zudem werden organisationsspezifische Kriterien hinsichtlich der Produktivität (Motivation, Effizienz, Effektivität, Aufgabenoptimierung, (zeitliche) Kosten) betrachtet.

Damit sollen die Forschungsfrage und vor allem die nachfolgend aufgeführten, sehr subjektiv wahrgenommenen Anforderungen evaluiert werden:

- FR1: Übersichtliche Darstellung des Empfehlungsergebnisses
- FR2: Nachvollziehbare Begründung der Empfehlungsentscheidung anhand der Eingaben
- FR8: Förderung der Bereitschaft der Anwendung von XAI
- NFR3: Benutzerfreundlichkeit

## 6.2. Ergebnisse der Nutzerevaluation

Nachfolgend werden die Ergebnisse der Nutzerevaluation zusammengefasst und es wird auf besondere Aspekte eingegangen, die in Verbindung mit den subjektiven, zu evaluierenden Anforderungen stehen.

### Nutzen

Der erste Eindruck des XAIRs hinsichtlich der Benutzerfreundlichkeit (NFR3) war durchweg positiv und die Personen der unterschiedlichen Zielgruppen benötigten keine externe Hilfe für seine Anwendung. Sie empfanden ihn als überraschend einfach zu benutzen (Person A), vor allem in Relation zu der Aufgabe, für die man ihn nutzt (Person D). Person B hatte ihn komplexer erwartet und empfand ihn als „leichtgewichtig in der Anwendung“.

Der Aufbau des XAIRs wurde generell als intuitiv und ansprechend wahrgenommen. Die Schritt-für-Schritt Durchführung anhand der Navigationsleiste wurde von vier der fünf Befragten als intuitiv und nutzerfreundlich bezeichnet. Allerdings wurde die Navigation von einer Person nicht sofort gesehen (Person D) und von einer erst während des Interviews bemerkt (Person B), weshalb letzterer der Zugang zu den Detailseiten innerhalb der Zeit des eigenständigen Ausprobierens verwehrt war.

Die bereitgestellten, textuellen Hilfestellungen wurden vor allem bei der Eingabe der benötigten Parameter von allen genutzt und gebraucht. Durch ihre Verwendung kamen alle Befragten sehr gut mit den verwendeten Termini für den Empfehlungserhalt zurecht. Person C ist wie Person E und B der Meinung, dass aufgrund dieser keine tieferen ML-/XAI-Kenntnisse notwendig sind. Allerdings glaubt sie, dass wenn es um ein „bisschen mehr als oberflächlich ‚Was kann die Methode?‘ geht“, dieses doch erforderlich wird. Person D meint, für die Nutzung „muss [man] natürlich [ein] grundlegendes Verständnis von Statistik haben, um das zu verwenden. Aber jetzt speziell ML muss man nicht können“. Laut Person A „braucht [man] schon eine Grundausbildung“ bzgl. XAI, bspw. in Form eines Vortrags, da Begriffe wie lokal und global für den durchschnittlichen Data Scientist, der noch nicht mit XAI gearbeitet hat, unbekannt sind.

Die Personen außerhalb des Data Science Kontexts empfanden zudem den einleitenden, generellen Text über XAI als einen guten Einstieg in die Thematik mit der Möglichkeit, sich darüber genauer zu informieren.

### Ergebnisqualität

Alle Interviewteilnehmer gaben an, dass die Informationen der eigentlichen Empfehlung und ihre Begründung für sie am interessantesten sind, und sie diese schnell gefunden haben. Person C war zudem besonders an der Liste der Fragestellungen interessiert, die man sich für die Auswahl einer XAI-Methode vergegenwärtigen muss.

Die Ergebnisseite und die Darstellung der Methodeneignung als Säulendiagramm wurde generell als gut und übersichtlich bewertet (FR1). Da die Säulen mit den Methodenabkürzungen beschriftet sind und diese den Personen mit wenigen Methodenvorkenntnissen unbekannt waren, reagierten sie zunächst

verwirrt. Außerdem war teilweise ein mehrmaliges Lesen der Hover-Texte der Säulen erforderlich, welche die von der XAI-Methode beantworteten Fragen aufführen.

Die Entscheidung bzgl. der Skalierung der Säulen fiel während der Konzeption aufgrund vieler verschiedener Darstellungsmöglichkeiten und Interpretationsoptionen schwer (vgl. 4.3.6). Sie wurde von niemandem explizit angesprochen oder als irritierend wahrgenommen. Person A „fand es besonders schön, dass die Ergebnisse, die [sie] bekommen ha[t], sehr distinguiert waren“.

Ohne explizite Nachfrage wurde die Nicht-Volatilität des Empfehlungssystems einmal gelobt (Person A) und einmal aufgrund der unwesentlichen Eignungsänderungen als irritierend empfunden (Person B). Person B änderte dabei die Ausprägung von Eingabe(n) durch eine Anpassung der Position eines Fuzzy-Schiebereglers von „Mitte bis Vollausschlag“, was nur die Änderung der Reihenfolge der Methoden 2-5 zur Folge hatte.

Die tabellarische Begründung der Empfehlungsentscheidung und der Eignungsvergleich der XAI-Methoden anhand der Eingabeparameter (FR2) wurden von allen als sehr positiv und übersichtlich wahrgenommen. Person D sieht „einfach auf einen Blick, für was ist eigentlich welche Methode wie geeignet“ und für Person A ist diese Tabelle „interessant, da [sie] gemerkt ha[t]: Dahinter steht tatsächlich eine systematische Entscheidung“. Letztere findet den dadurch ermöglichten Vergleich hinsichtlich der Eingaben sehr aufschlussreich, da dadurch die Frage „Warum Methode X und nicht Y?“ beantwortet wird. Solche, von Menschen verlangten, kontrastiven Erklärungen werden als sehr intuitiv und wertvoll eingeschätzt (Miller 2019, S. 20).

Aufgrund der tabellarischen Erklärung ist die Entscheidung für alle Personen nachvollziehbar, was wichtig ist, um dem System vertrauen zu können. Allerdings würde

- Person A vor der produktiven Anwendung der empfohlenen XAI-Methode eine eigene, genaue Recherche bzgl. dieser durchführen.
- Person B im Zweifel eher auf die Empfehlungen ihres Teams vertrauen, obwohl sie die dargestellte Entscheidungsbegründung für ausreichend empfindet, um dem System zu vertrauen.
- Person C die Empfehlung annehmen, aber sich nicht nur die erste, sondern mehrere der höchst empfohlenen Methoden genauer ansehen. Sollte nichts konkret gegen die Anwendung einer bereits bekannten XAI-Methode sprechen, die ihre Fragestellungen gut beantwortet, würde sie diese wahrscheinlich auch nochmals austesten.
- Person D einem System nie blind vertrauen, aber dem Rat des XAIRs folgen und zunächst die empfohlenen Methoden betrachten.

Nur Person E empfindet die Ergebnisse des XAIRs aufgrund ihres mangelnden XAI-Fachwissens als „nicht vertrauenswürdig“. Das System könnte ihr Vertrauen nur durch die Angabe eines konkreten Eignungsgrades der XAI-Methoden, bspw. durch eine Prozentzahl, erhöhen.

Obwohl sich keiner der Interviewteilnehmer im Detail mit den Methoden- und Implementierungsinformationen auseinandergesetzt hat, ist „die Aufbereitung [...] so, dass [Person B] auf jeden Fall

reingucken würde, weil es sich recht bekömmlich für [sie] anfühlt.“ Die Inhalte sind, soweit von den Personen beurteilbar, in kurze Abschnitte gegliedert und verständlich aufbereitet. Vor allem die als Fragen formulierten Überschriften wurden von allen als positiv wahrgenommen. Für Person D ist dies „für den Einstieg eine sehr gute Menge an Information“, es macht „den Eindruck, als dass es sich lohnt, diese Info durchzulesen. Weil es ist nicht zu viel, aber es ist schon sehr gezielte Information [...]“. Besonders die Antwort auf „Welche Frage beantwortet die XAI-Methode *nicht*“ ist laut ihr „Gold wert“. Sollte Person A alles lesen müssen, empfände sie die Seiteninhalte als zu viel Information. Sie schätzt daher das Trichterprinzip der Aufbereitung, d.h. dass die Wichtigkeit der aufgeführten Information mit Voranschreiten des Textes abnimmt und der Detailgrad zunimmt. Dadurch kann sie „einfach irgendwo aufhören zu lesen und ha[t] nicht das Gefühl, [sie] ha[t] jetzt was Wichtiges verpasst“.

## Produktivität

Der zeitliche Aufwand der Nutzung des XAIRs war für alle Befragten gering und akzeptabel. Da sich keiner näher mit den Methoden- und Implementierungsdetailseiten auseinandergesetzt hat, kann die Zeitinvestition zum Verstehen des Empfehlungsergebnisses nicht beurteilt werden.

Die Zeitersparnis durch die Benutzung des XAIRs bei der Methodenauswahl wird vor allem von den Personen A und D als sehr groß empfunden und besonders geschätzt. Dies kann damit begründet werden, dass Person A mit der rasanten Entwicklung und Zerstreutheit des Bereichs der XAI vertraut ist. Auch Person D „weiß, was es bedeutet, sich so Wissen erstmal zu erarbeiten und deswegen kann [Person D] ganz gut einschätzen, wieviel Zeit [...] das hier spart“. Nach eigener Aussage „extrem viel Zeit“.

Ein weiterer Aspekt, auf welchen nachfolgend näher eingegangen wird, ist die Eingabe der benötigten Parameter (FR3), deren Umfang (vgl. FR3.4) nicht explizit zur Sprache kam.

Die Personen A und D hatten keine Probleme, die Eingabeparameter aus den aufbereiteten Daten des Anwendungsfalls zu lesen; für Person A kam nur eine Kombination der Eingaben infrage.

Den anderen Befragten fiel die Ermittlung der datenspezifischen Eingaben teilweise schwerer: Für Person B war unklar, wie weit sie den Ausschlag des Fuzzy-Schiebereglers ziehen soll und auch C und E hatten durch Fehlen eines „Ankerpunktes“ bzw. eines Auswahlbereichs Probleme, die Eingaben in Relation zu setzen. Person D hat sich zudem gefragt, inwiefern sich eine exakte Eingabe im Vergleich zur ungefähren auf die Empfehlungsergebnisse auswirkt und ob sich die genaue Ermittlung dieser lohnt. Sie versteht, warum die Web-Anwendung aus Datenschutzgründen eine manuelle, ungefähre Eingabe erlaubt, bewertet eine subjektive Einschätzung der Daten allerdings als kritisch, da dadurch eine Voreingenommenheit des Nutzers bzgl. der Daten für die Auswahl einer XAI-Methode berücksichtigt wird.

Die aktuell in der GUI nur angedeutete Funktionalität einer automatisierten Datenanalyse würden die Personen A und C sehr begrüßen; Person D würde diese „sofort benutzen und auch definitiv irgendeinem System vorziehen, wo [sie] irgendwelche gefühlten Korrelationen oder so eingeben muss.“ Sie würde gerne jede weitere Verzerrung der Daten minimieren, die durch eine ungefähre Einschätzung dieser Werte entstehen könnte.

Person C war bei der Parametereingabe außerdem nicht klar, ob die Angabe der Präferenz eines Erklärungsumfangs, z.B. „lokal“, als Ausschlusskriterium für globale XAI-Methoden gilt.

Weitere Informationen, wie (z.B. in Kombination mit welchen anderen Kriterien) und warum sich die Eingabe auf die Methodeneignung auswirkt, wurden zudem von den Personen A und C gewünscht. Dies kam ein Mal bei der Eingabemaske und ein Mal bei der erklärenden Tabelle zur Sprache.

## 6.3. Diskussion der Resultate

In diesem Kapitel werden die Ergebnisse aus der Nutzerevaluation diskutiert und interpretiert. Zunächst wird die Nutzerzufriedenheit anhand der Erfüllung ihrer initialen Erwartungshaltungen beurteilt. Anschließend wird eine Antwort auf die Forschungsfrage aus den Ergebnissen der Evaluation abgeleitet und der prototypische XAIR in den aktuellen Forschungsstand eingeordnet.

Zum Schluss wird auf die Limitationen der Evaluation und die damit identifizierten, zukünftige Herausforderungen eingegangen.

### 6.3.1. Erfüllung der Erwartungshaltung der Nutzer

Die Personen ohne Data Science Kontext (Personen B und D) haben den XAIR als ein einfach zu bedienendes und selbsterklärendes Tool erwartet. Person B hat die gleichen Erwartungen wie an einen menschlichen Experten und erhofft sich den Erhalt einer Empfehlung ohne technische Expertise; Person E erwartete, dass für die resultierende Empfehlung keine weiteren Quellen herangezogen werden müssen.

Die Erwartungshaltung von Person A mit XAI-Erfahrung war, dass der XAIR eine „Wissensbasis [verwendet,] die größer ist als das, was [Person A] als Data Scientist [sich] selber realistisch aufbauen kann“. Dadurch hofft sie, aktuelle und unbekannte XAI-Verfahren ohne das Durcharbeiten des dazugehörigen Papers kennen zu lernen und die Tendenz zu überwinden, nur bereits bekannte, eventuell ungeeignete Verfahren anzuwenden. Ihr Ziel ist es, durch die Nutzung eine innerliche Sortierung durch Kategorisierung der Methoden zu schaffen und bei der tatsächlichen Auswahl Zeit zu sparen.

Auch Person C hofft auf den Erhalt eines Überblicks über die aktuellen XAI-Methoden, ihre Ergebnisse und alle zu berücksichtigenden Aspekte. Person D erwartet einen Methodenvergleich anhand des konkreten Anwendungsfalls und ihrer persönlichen Präferenzen.

Grundsätzlich wurden die Erwartungen aller Teilnehmer an das XAI-Empfehlungssystem maßgeblich erfüllt und die von Person D, durch die Bereitstellung eines Implementierungsvorschlags mit Hinweisen, sogar übertroffen. Sie empfindet das als „super coole[n] Service, [...] unabhängig von der [...] eigentlichen Empfehlung“ und würde daher den XAIR nochmals aufrufen, auch wenn sie bzgl. der Methodenauswahl schon recht sicher wäre.

Alle Personen würden den XAIR nochmals bzw. mehrmals nutzen. Person C würde ihn allerdings nur erneut befragen, wenn das zu erklärende Modell sich während der Entwicklung stark verändert, oder ein neues Modell eines anderen Kontexts erklärt werden soll. Person B hat nicht den Anspruch, den XAIR für einen Empfehlungserhalt zu nutzen und gleichzeitig alle Methoden zu erlernen. Trotzdem traut sie sich bei einer häufigeren Verwendung durch die dadurch gesammelte Erfahrung zu, eigenständig eine geeignete XAI-Methode für vergleichbare Modelle auszusuchen. Person A würde Anderen seine Nutzung nicht nur empfehlen, sondern sie dazu „sogar zwingen, [...] damit sie endlich mal ihre

XAI-Methoden anwenden und nicht immer die Ausrede benutzen, ja, aber die Accuracy ist doch schon bei 95%, was muss ich dann noch erklären“.

Die Anwendung des XAIRs ist Voraussetzung, um bei der Auswahl und Anwendung von XAI-Methoden unterstützende Empfehlungen geben zu können. Die Motivation seiner Wiederverwendung deutet auf die Zufriedenheit der Nutzer mit den daraus gewonnenen Empfehlungsergebnissen und ggf. Erkenntnissen hin. Der XAIR kann ihm die Verwendung und das Verstehen der XAI-Methoden zwar nicht abnehmen, allerdings kann er laut Person A als „Treibmittel“ angesehen werden, um Unternehmen zur Auseinandersetzung mit XAI zu zwingen und Data Science Best-Practices einzuhalten.

Hinsichtlich der Motivation zur anschließenden Anwendung der empfohlenen XAI-Methode(n) äußerte sich Person B nicht, da sie das konstruierte Szenario als zu hypothetisch empfand. Person D würde sich vor einer Anwendung zuerst eine Empfehlung ihres Data Science Teams einholen, da ihr XAI Hintergrundwissen fehlt. Person A würde nach Absprache mit dem Kunden, dem Adressaten der Erklärung, die Empfehlung anwenden und auch Person C würde mehrere dieser ausprobieren wollen. Person D, welche in der Forschung primär mit Modellen basierend auf Bilddaten arbeitet, äußerte sich überraschenderweise mit „Ja. Tatsächlich bin ich wirklich motiviert, auch wenn ich es eigentlich gar nicht brauche“.

### 6.3.2. Erkenntnisse bezüglich der Forschungsfrage

Das resultierende XAI-Empfehlungssystem unterstützt den Nutzer, indem es die subjektiven Anforderungen und Systemziele erfüllt. Die qualitative Nutzerevaluation gibt Aufschluss darüber, inwieweit und aus welchen Gründen der XAIR zur Erreichung dieser Ziele beiträgt und wie die Forschungsfrage beantwortet werden kann.

Zusammenfassend weist sie darauf hin, dass der XAIR den Nutzer durch Empfehlungen unterstützt, indem diese

- ihm für den Daten-, Modell- und Nutzungskontext angemessene XAI-Methoden liefern
- ihm die für die Auswahl zu berücksichtigende Aspekte aufzeigen und ihn somit diesbezüglich sensibilisieren
- ihm Zeit bei der Informationsbeschaffung sparen
- nachvollziehbar erklärt werden
- untereinander vergleichbar sind
- ihm eine Rückversicherung geben, dass er mit einer konkreten XAI-Methode auf dem richtigen Weg ist
- aktuelle XAI-Methoden bei der Auswahl berücksichtigen, sodass dem Nutzer ein Überblick über den aktuellen XAI-Forschungsstand gegeben wird und er zur häufigeren Anwendung des XAIRs motiviert ist

Die Evaluation gibt Aufschluss darüber, dass der Nutzer, sowohl für die Auswahl als auch die Anwendung, zusätzlich durch die Aufbereitung von strukturiertem, tieferen Methodenwissen und Implementierungshinweisen unterstützt werden kann. Die empfehlungsunabhängige Möglichkeit, ohne Zuhilfenahme (aber mit Angabe) weiterer Quellen tiefere Methodeneinblicke zu erhalten, wurde sehr wertgeschätzt.

### 6.3.3. Einordnung des Ergebnisses in den Stand der Forschung

Der Orientierungsbaum von Kraus et al. (2021) verfolgt ähnliche Ziele wie der XAIR und gibt konkrete Methodenempfehlungen anhand von Eigenschaften der Zielgruppe und des Daten- und Modelltyps (siehe Kapitel 2.2.2). Den aktuellen Forschungsstand überblickend, ist dieser mit dem XAIR am vergleichbarsten.

Der Orientierungsbaum beschränkt sich nicht nur auf tabellarische Daten, weshalb die Auswahl an XAI-Methoden, die sich für solche eignen, kleiner ist. Die Studie von Kraus et al. (2021) bietet bzgl. der aufgeführten Methoden auch einige Informationen und ein konkretes Anwendungsbeispiel. Diese Informationen sind übersichtlich strukturiert in „Art der Erklärung“, „Anwendbar auf“, „Technischer Hintergrund“ und „Vor- und Nachteile“. Das Paper bietet daher einen zeitsparenden Einstieg bei der Auseinandersetzung mit der XAI-Methode und eine gute Orientierung, falls ein Nutzer eine Rückversicherung für seine bereits ausgewählte Methode sucht.

Der XAIR ist im Gegensatz dazu umfassender, bietet bspw. auch Implementierungsempfehlungen und somit eine konkrete Unterstützung bei der Methodenanwendung.

Der statische Orientierungsbaum beurteilt die Methodeneignung primär nach der Zielgruppe, wobei Wissen subjektiv und auch innerhalb einer Zielgruppe individuell ist. Einem KI-Entwickler ohne Erfahrung mit der Verwendung von XAI empfiehlt er z.B. standardmäßig dieselbe Methode.

Der XAIR berücksichtigt bei seiner Empfehlung mehr eignungsbeeinflussende Kriterien als die Orientierungshilfe von Kraus et al. (2021). Seine Empfehlungen basieren nicht hauptsächlich auf der Erfahrung des Nutzers; vielmehr soll er dem Nutzer durch eine übersichtliche, ausführlichere Aufbereitung des Methodenwissens auch neue (aktuelle) XAI-Methoden näherbringen.

Außerdem ist der XAIR, im Gegensatz zum Orientierungsbaum, hinsichtlich Kriterien und XAI-Methoden einfach erweiterbar.

Der wohl größte Vorteil des XAIRs gegenüber dem Orientierungsbaum ist, dass er seinen Vorschlag geeigneter XAI-Methoden hinreichend begründen kann. Die Evaluation weist darauf hin, dass diese Begründung neben der eigentlichen Empfehlung für alle Nutzer am interessantesten ist. Eine Nachvollziehbarkeit der empfohlenen Methodenauswahl ist durch den XAIR gegeben und die Auswirkung jedes Kriteriums des Nutzungskontexts klar erkennbar. Die Eignungen der verschiedenen XAI-Methoden ist somit einfach vergleichbar.

### 6.3.4. Limitationen und Herausforderungen

#### Einschränkung des Projektumfangs

Da der gesamte Entwicklungsprozess im Rahmen des Requirement Driven DSR Ansatzes auf den initialen Anforderungen basiert, können sich diese währenddessen durch Hinzufügen, Spezifizieren oder Revidieren verändern. Gründe dafür können ein Mangel an geeigneten Umsetzungsmethoden, eine zu generische Formulierung der Anforderung oder ein Wissenszuwachs sein, der entweder während der Entwicklung oder aufgrund neuer oder vergleichbarer Ansätze des Forschungsstandes entstand. (Braun et al. 2015, S. 11)

Auch die Anforderungen an den XAIR wurden während der Entwicklung angepasst bzw. reduziert. Die Anforderung der Angabe eines spezifischen Erklärungsformates (FR4.2) wurde während der Konzeption auf die Angabe eines gewünschten Erklärungsumfangs reduziert und somit generalisiert. Dadurch werden spärliche oder leere Empfehlungsergebnismengen der anwendbaren Methoden aufgrund einer zu geringen initialen XAI-Methodenunterstützung vermieden. Auch die Angabe der Komplexität/Zielgruppe der Erklärung (FR4.3) wurde, aufgrund mangelnder Beurteilbarkeit trotz Kenntnisnahme des aktuellen Forschungsstandes, verworfen.

Die Speicherbarkeit der Empfehlungsergebnisse (FR6) wurde nachträglich aus dem Umfang entfernt, da die Nachfrage der Nutzer danach nicht hoch ist und der erneute Erhalt eines spezifischen Empfehlungsergebnisses durch Eingabe weniger Parameter möglich ist. Von NFR4 wurde die Möglichkeit des Hinzufügens neuer Expertenmeinungen für die Methodenbewertungen nicht umgesetzt. Eine Aggregation des Expertenwissens durch die Aufnahme neuen Wissens führt zu Informationsverlust, da die Einschätzung der Methodeneignung gemittelt werden muss, sofern sich die Meinungen der Experten hinsichtlich dieser unterscheiden. Daher wurde auf die Anforderung bei der Umsetzung zunächst verzichtet. Aufgrund eines zusätzlichen Mangels weiterer Expertenmeinungen und geeigneter Aggregationsmethoden findet die Methodenbewertung aktuell, wie bereits erwähnt, nur durch die Autorin statt. Durch dieses nicht-automatisierte Editieren der Wissensbasis sollen XAI-Experten außerdem zum bewussten, gemeinschaftlichen Austausch angeregt werden.

#### Limitationen der Evaluation

Die Umsetzung der Portabilität des XAIRs (NFR1) wurde nicht explizit evaluiert. Für die effektive Verwendung des XAIRs innerhalb einer ML-Pipeline bedarf es einer weiteren Datenanalyse-Komponente zur Ermittlung der datenbezogenen Eigenschaften. Dafür wurde bereits ein Konzept ausgearbeitet, welches allerdings nicht umgesetzt wird, um den Rahmen dieser Arbeit nicht zu sprengen.

Die gewünschte Portabilität der Anwendung ist dennoch gegeben. Die Projekte sind mit Docker containerisiert und somit, dank der Kapselung der Anwendungen und ihrer Abhängigkeiten, als Microservices in einer beliebigen Umgebung einsetzbar. Der XAIR ist somit minimal-invasiv in eine Komponente einer Kubeflow Pipeline integrierbar. In der ML-Pipeline müssen ihm lediglich die zuvor ermittelten oder in der Konfiguration festgelegten Kriterienwerte in Form von *inputValues* oder einem *inputPath* übergeben werden (vgl. Abbildung 4.4, S. 42).



Eine weitere Limitation ist, dass das Feedback dieser abschließenden Evaluation nicht umgesetzt werden kann, da die daraus gewonnenen Vorschläge im Rahmen dieser Arbeit nicht weiter evaluierbar sind.

Kleinere Änderungen, die nicht im Konflikt mit Aussagen anderer Nutzer stehen, wurden allerdings vorgenommen: Auf der Ergebnisseite wird nun explizit auf die Navigationsleiste hingewiesen und dem Nutzer die Möglichkeit einer alternativen Navigation mittels Links gegeben. Außerdem werden die Hover-Texte des Ergebnis-Säulendiagramms ausformuliert, um einer anfänglichen Verwirrung der XAI-Fachfremden entgegenzuwirken. Sie werden mit dem vollen Methodennamen und einem „beantwortet die Frage“ eingeleitet, bevor die von der jeweiligen XAI-Methode gegebenen Antworten aufgeführt werden. Es wurden keine Veränderungen bei Funktionen vorgenommen, die generell als positiv empfunden wurden (z.B. an den Hilfetexten der Eingabeparameter oder der erklärenden Tabelle).

Zudem ist eine Ausweitung der Nutzerstudie denkbar, da diese aus Zeitgründen nur auf fünf Personen und auf das Einholen einer Methodenempfehlung limitiert ist. Auf Grundlage dieser Arbeit wäre auch eine längere Studie möglich, die die Umsetzbarkeit der Empfehlung mit den gegebenen Informationen und die situationsspezifische Qualität der resultierenden Erklärung untersucht. Sie ist notwendig, um zu beurteilen, ob der XAIR den Nutzer bei der Methodenanwendung tatsächlich unterstützt. Bzgl. der Qualitätsbeurteilung des Methodenergebnisses ist sie allerdings zum jetzigen Zeitpunkt schwer möglich, da sie abhängig von den Erkenntnissen weiterer Forschungsfelder ist, welche sich mit Erklärungen und deren (subjektiver) Wahrnehmung und Qualitätsbeurteilung auseinandersetzen (vgl. Kapitel 4.1.1).

### **Zielgruppendefinition**

Person B in der Rolle eines für das Modell verantwortlichen Product Owners stellte sich die Frage, ob sie überhaupt Teil der Zielgruppe des Systems ist oder ob der XAIR nicht eher von den Data Science Teams selbst genutzt werden sollte. „Für [sie] fühlte sich das an wie eine extra Schleife“: Sie erhielt Hilfe des Data Science Teams bei der Parametereingabe (durch anbei liegende Informationen des Anwendungsfalls) und bekam ein Ergebnis, welches das Team ihr laut eigener Aussage wahrscheinlich selbst hätte geben können. Allerdings „weiß [sie] aber nicht, ob jemand aus dem Team diese Antwort geben könnte“. Das wirft die Frage auf, ob der Umfang der Zielgruppe auf Modellentwickler und ML-/XAI-Interessierte reduziert werden sollte, was eine Änderung der Anforderungen bewirken kann.

Unabhängig von dem XAI-Wissen seines Data Science Teams kann ein Product Owner mithilfe des XAIRs die Initiative ergreifen und ggf. den Einsatz von XAI für das ML-Modell fordern und somit generell fördern. Abgesehen von der Frage, von wem der XAIR konkret genutzt werden sollte, ist der Modellverantwortliche als Konsument der Erklärung für die Auswahl einer geeigneten XAI-Methode ebenfalls sehr wichtig. Da die Ergebnispräsentation außerdem sowohl für die ML-Fachfremden, als auch für die potenziell einem Data Science Team zugehörigen Personen A, C und D ansprechend war, wird die Zielgruppe des XAIRs im Nachhinein nicht angepasst. Dadurch wird auch Personen mit wenigen ML-/XAI-Kenntnissen die Möglichkeit gegeben, einen einfachen Einstieg in diese komplexe und wichtige Thematik und in konkrete XAI-Methoden zu erhalten.

### **Zukünftige Herausforderungen**

Obwohl der XAIR, wie oben beschrieben, den Nutzer zu seiner Anwendung zu motivieren und ihn durch Empfehlungen zu unterstützen scheint, gibt es einige zukünftige Herausforderungen, die während der Evaluation deutlich wurden.

Eine zukünftige Herausforderung stellt die Behebung der Schwierigkeiten bei der Eingabe der benötigten Parameter dar. Es muss die Frage beantwortet werden, wie der Nutzer dabei unterstützt werden kann, weitestgehend unabhängig von Data Science Fachwissen eine Einschätzung der vagen Eingabeparameter vorzunehmen. Eine mögliche Lösung bietet die Umsetzung des Konzepts der automatisierten Datenanalyse (vgl. Kapitel 4.3.2). Sie würde von den Befragten gerne genutzt werden und reduziert zudem die von Person D erwähnte, mögliche Voreingenommenheit des Nutzers bzgl. der Daten.

Es sollte außerdem geklärt werden, welche Fragen (mindestens) noch beantwortet werden müssen, um den unterschiedlichen Erklärungsbedürfnissen der Nutzer gerecht zu werden. Die Evaluation gibt Aufschluss darüber, dass teilweise weitere Fragen bzgl. der Auswirkungen der Eingabekombinationen auf die XAI-Methodeneignungen bestehen. Ihre Beantwortung setzt allerdings die Beachtung diverser Aspekte der Nutzbarkeit (Usability) voraus. Manche Nutzer vermissen diese Informationen nicht und sie könnten, abhängig von der Darstellung und dem Fachwissen des Nutzers, als zu komplex oder erschlagend empfunden werden (vgl. die Resonanz bei der Ergebnisdarstellung bzgl. den aufgeführten Fragen, die von XAI-Methode beantwortet werden).

Obwohl für Person E die Steigerung des Vertrauens nicht durch Verbesserung des XAIRs, sondern ausschließlich über den Aufbau eigenen Fachwissens möglich ist, könnten ausführlichere Erklärungen dabei helfen. Die von ihr gewünschte, prozentuale Angabe der Eignung ist allerdings nicht möglich.

Das System legt zwar sein ganzes Regelwerk transparent offen, allerdings sollte eine gewisse Skepsis bzgl. der sehr vom Anwendungskontext abhängigen Empfehlungen des XAIRs beibehalten werden, zumal dessen Wissen ausschließlich die Meinung einer Expertin widerspiegelt. Vor allem im Bereich wissensbasierter Systeme haben Fuzzy-Lösungen außerdem einen heuristischen Charakter, weshalb Optimalität und Stabilität nicht gewährleistet sind (Nissen 2007, S. 25). Daher ist die Nachprüfung einer XAI-Methodenempfehlung, z.B. in Form einer anschließenden Recherche oder einem Vergleich diverser Verfahren, sogar erwünscht.

## 7. Fazit und Ausblick

Nachfolgend werden die Ergebnisse der Arbeit in einem Fazit zusammengefasst und ein Ausblick auf zukünftige Erweiterungs- und Verbesserungsmöglichkeiten des entstandenen XAI-Empfehlungssystems gegeben.

### 7.1. Fazit

Im Rahmen dieser Arbeit wird gezeigt, wie man den Nutzer bei der Auswahl und anschließenden Anwendung von XAI-Methoden auf Black-Box Modelle durch Empfehlungen unterstützen kann. Dafür wird nach dem Requirements Driven DSR Ansatz ein XAI-Empfehlungssystem (XAIR, XAI Recommender) konzipiert und implementiert, welches dem Nutzer ohne tiefe ML-Kenntnisse für den Anwendungskontext geeignete XAI-Methoden vorschlägt.

Die Eignung einer XAI-Methode lässt sich nicht aufgrund der Qualität der resultierenden Erklärung beurteilen, sondern anhand von Eigenschaften des Modell-, Daten- und Nutzungskontexts. Diese machen die Anwendung einer XAI-Methode entweder unmöglich (Ausschlusskriterien) oder haben einen negativen, verfälschenden Einfluss auf die resultierende Erklärung oder ihre Interpretierbarkeit. Die eignungsbeeinflussenden Kriterien haben eigenständig und/oder in Kombination mit anderen eine (meist negative) Auswirkung auf die Anwendbarkeit einer XAI-Methode.

Konkrete Kriterien, wie bspw. der Korrelationsgrad eines Datensatzes, müssen aus den im XAIR integrierten XAI-Methoden abgeleitet werden. Das ist auch bei der Erweiterung des XAIRs zu berücksichtigen. Für die initiale Methodenauswahl werden die in den Richtlinien von Kangur (2020) und Belle & Papantonis (2020) zur Anwendung empfohlen Verfahrensarten, sowie die laut Bhatt et al. (2020) häufig im Deployment eingesetzten *Counterfactual Explanations* berücksichtigt. Als Visualisierungstechniken sind *Partial Dependence Plot (PDP)* in Kombination mit *Individual Conditional Expectation (ICE)*, und *Accumulated Local Effects (ALE)* vertreten. *Anchors* und *Counterfactuals guided by Prototypes (CFProto)* werden als lokale Methoden, und *Permutation Feature Importance (PFI)* als globales Verfahren hinzugefügt. Die Auswahl wird um *SHapley Additive exPlanations (SHAP)* ergänzt, die sowohl lokale, als auch globale Erklärungen liefern.

Die in einer Literaturrecherche und einer Anforderungsanalyse ermittelten Kriterien dienen dem XAIR als Eingabeparameter. Für eine Empfehlung kann das System entweder über eine Web-Anwendung, oder direkt bzw. automatisiert durch Integration in eine ML-Pipeline konsultiert werden.

Das zweistufig operierende System bestimmt zunächst mithilfe eines Fuzzy-Expertensystems die Methodeneignung anhand der eignungsreduzierenden Kriterien. Die darin angewandte Fuzzy-Logik adressiert die diversen Probleme der Ungenauigkeit, die im Kontext von XAI auftreten: Sie vermeidet die Definition von Schwellwerten der eignungsbeeinflussenden Kriterien, welche für eine exakte Einschätzung der Eingaben notwendig wären. Außerdem ermöglicht sie die qualitative Modellierung des ungenau-

en, in der Literatur zu findenden Expertenwissens und erlaubt eine Vagheit und Nicht-Eindeutigkeit möglicher Empfehlungsergebnisse.

Für die Bestimmung der Methodeneignungen sind die Auswirkungen der Kriterien auf die XAI-Methoden identifiziert und in Fuzzy-Regeln der Wissensbasis des Fuzzy-Expertensystems formalisiert. Um einen konkreten Eignungswert aus den Ergebnissen der aktivierten Regeln zu erhalten, wird die COG-Defuzzifizierungsstrategie angewendet.

Aus der Menge aller XAI-Methoden eliminiert der XAIR im zweiten Schritt diejenigen, welche durch die Ausschlusskriterien nicht anwendbar sind.

Bei der Nutzung der Web-Anwendung werden dem Nutzer im Anschluss die einsetzbaren XAI-Methoden übersichtlich und nach ihrer Eignung sortiert präsentiert.

Eine Evaluation qualitativer Nutzerinterviews gibt Aufschluss darüber, dass sich die Nutzer durch die konkreten, begründeten und vergleichbaren Empfehlungen des webbasierten XAIRs bei der Auswahl unterstützt fühlen. Der XAIR bietet zudem Unterstützung bei der Anwendung der XAI-Methoden, indem er durch strukturierte Bereitstellung tieferen Methodenwissens und konkreten Implementierungshinweisen die Komplexität der Informationsbeschaffung bzgl. XAI reduziert. Er dient als Nachschlagewerk eingepflegter XAI-Methoden und stellt generelle, bewährte Vorgehensweisen zur Förderung der Nachvollziehbarkeit von ML-Modellen bereit.

Die Evaluation weist darauf hin, dass der Nutzer durch die Zeitersparnis dieser Empfehlungen nicht nur bei der Methodenauswahl unterstützt wird, sondern auch zur erneuten Konsultation des XAIRs und zur Anwendung der XAI-Methode motiviert ist.

Diese Arbeit trägt zum Stand der Forschung bei, indem sie eine, in dieser Ausführlichkeit noch nicht vorhandene und erweiterbare Liste konkreter Kriterien ermittelt, welche die Eignung einer XAI-Methode beeinflussen. Durch die Aggregation von Wissen verstreuter Literatur identifiziert und formalisiert sie außerdem die Auswirkungen dieser Kriterien auf die Eignung ausgewählter XAI-Methoden.

Der aus dieser Arbeit resultierende XAIR liefert schnell Empfehlungen für XAI-Methoden, die sich für den spezifischen Anwendungskontext eignen. Als erstes Empfehlungssystem im Kontext von XAI stellt er dieses Wissen den Nutzern ohne tiefe ML-Kenntnisse strukturiert, in einer nutzerfreundlichen Web-Anwendung bereit. Durch eine Begründung der Empfehlungsgenerierung, einer Vergleichbarkeit der verschiedenen XAI-Methodeneignungen und einer Zeitersparnis bzgl. der XAI-Informationsbeschaffung wird der Nutzer erfolgreich bei der Auswahl einer XAI-Methode unterstützt. Außerdem wird er für die Eigenschaften des Anwendungskontexts sensibilisiert, die es bei der Auswahl zu beachten gilt. Durch Empfehlungen konkreter Implementierungen und technischer und methodischer Aspekte wird außerdem die Bereitschaft der anschließenden Anwendung der XAI-Methode(n) erhöht.

Durch seine umfassenden und nachvollziehbaren Empfehlungen, seine nutzerfreundliche Anwendung und seine Erweiterbarkeit, hebt der XAIR sich im Besonderen von aktuellen Alternativen wie dem statischen Orientierungsbaum von Kraus et al. (2021) ab, der unbegründet Empfehlungen für XAI-Methoden auf Basis weniger Kriterien liefert.

## 7.2. Ausblick

Der konzipierte und prototypisch implementierte XAIR bildet eine Grundlage für diverse zukünftige Weiterentwicklungen.

Mögliche funktionale Erweiterungen sind die Implementierung der erhobenen, allerdings noch nicht umgesetzten Anforderungen der Speicherbarkeit der Empfehlungsergebnisse und der Integration von Nutzerfeedback bei der Empfehlungsgenerierung. Die laut Ziegler & Loepp (2019) von Nutzern gewünschte, interaktive Einflussnahme auf die Bewertung der XAI-Methoden lässt sich bspw. hinsichtlich des sehr subjektiven Vorbereitungsaufwands verwirklichen. Dessen Beurteilung erfordert kein tiefes XAI-Expertenwissen, weshalb sie einfach in die Wissensbasis aufgenommen und für zukünftige Empfehlungen berücksichtigt werden kann.

Der implementierte XAIR ist für die Aufnahme weiterer XAI-Methoden und eignungsbeeinflussender Kriterien über eine Wissenserwerbskomponente ausgelegt. Ihr Zuwachs ist zukünftig daher nicht nur möglich, sondern erwünscht. Für die Erweiterung des XAIRs um zusätzliches Expertenwissen wäre die Entwicklung eines Konzepts zur Berücksichtigung mehrerer Expertenmeinungen ohne Informationsverlust notwendig. Dabei muss berücksichtigt werden, dass bei stark unterschiedlichen Meinungen bzgl. der Kriterienauswirkungen die resultierende, aggregierte Methodenbeurteilung sinnhaft ist und sich z.B. nicht einfach mittelt.

Der Fokus der prototypischen Implementierung lag nicht auf der Einhaltung von Gestaltungsprinzipien oder Aspekten der Nutzerinteraktion zur Verbesserung der Nutzererfahrung. Für künftige Weiterentwicklungen ist es daher lohnenswert, Erkenntnisse anderer Forschungsbereiche zu berücksichtigen, wie bspw. der HCI, des User Interface Designs (UI Design) und der User Experience (UX).

Außerdem ergeben sich weiterführende Forschungsfragen, die die Empfehlungsqualität des XAIRs betreffen. Hinsichtlich der Nachvollziehbarkeit der Empfehlungsermittlung und dem Vertrauen in das System stellt sich bspw. die Frage, welche Informationen den Nutzern in welcher Form gegeben werden müssen, um ihre unterschiedlichen Erklärungsbedürfnisse zu erfüllen.

Zudem ist die Umsetzung des ausgearbeiteten Konzepts der Datenanalyse-Komponente möglich, die die datenspezifischen XAIR-Eingaben automatisiert ermittelt.

Die Implementierung ist notwendig, um eine Integration des XAIRs in eine ML-Pipeline und somit einen automatisierten Empfehlungserhalt zu ermöglichen. Dies wird angestrebt, um den Schritt der Auseinandersetzung mit möglichen XAI-Methoden obligatorisch zu machen und somit die (gesetzlichen) Anforderungen hinsichtlich der Nachvollziehbarkeit des ML-Modells durchsetzen können. Außerdem reduziert eine Integration dieser Funktionalität in die Web-Anwendung die bestehenden Schwierigkeiten des Nutzers bei ihrer manuellen Einschätzung. Dadurch wird auch eine eventuelle Verzerrung der datenbezogenen Eingabeparameter durch seine eventuelle Voreingenommenheit verhindert.

Durch diese Komponente können dem Nutzer zudem Hinweise bei Auffälligkeiten oder Besonderheiten der Daten gegeben werden. So wird er auf eventuelle Probleme des Preprocessings der Trainingsdaten aufmerksam gemacht, welche sich ggf. negativ auf das Modell ausgewirkt haben.

Denkbar ist außerdem eine Erweiterung der Funktionalität des XAIRs um konkrete Empfehlungen gewisser Methodenkombinationen, um stärkere, vollständigere Erklärungen zu erhalten. Allerdings werden laut Adadi & Berrada (2018) den Ansätzen, die das Potenzial einer solchen Kombination diskutieren, im Vergleich zur Entwicklung neuer XAI-Methoden aktuell wenig Aufmerksamkeit geschenkt. Die Berücksichtigung neuer Erkenntnisse dieses Forschungsgebietes ist für den XAIR für eine optimale Empfehlungsgenerierung sinnvoll.

Die bisher durchgeführten Evaluationen des Prototyps sind auf die Konsultation des XAIRs, d.h. das Einholen von XAI-Methodenempfehlungen, limitiert. Abhängig von den Erkenntnissen bzgl. der Qualitätsbeurteilung von Erklärungen ist, wie in Kapitel 6.3.4 erwähnt, eine Ausweitung der Nutzerstudie denkbar. Dadurch ließe sich die Qualität der resultierenden Methodenempfehlung und ihre Umsetzbarkeit mit gegebenen Information evaluieren.

Der XAIR wird als Open-Source Projekt unter der folgenden URL bereitgestellt:

`https://github.com/viadee/xair`

Dadurch sollen nicht nur die Weiterentwicklung des Systems durch eine gemeinschaftliche Zusammenarbeit gefördert werden, sondern auch XAI-Experten zu einem Austausch angeregt werden. Das Ziel der Veröffentlichung des XAIRs ist es, Personen für das Zugänglichmachen von anwendungsbefähigten Informationen zu XAI zu begeistern und generell Aufmerksamkeit und Bewusstsein für XAI und die Problematik von ML-Modellen zu generieren.

Zusammenfassend kann gesagt werden, dass das konzipierte XAI-Empfehlungssystem das Problem der fehlenden Unterstützung bei der Auswahl und Operationalisierung von XAI-Methoden adressiert und als „Treibmittel“ zur Auseinandersetzung mit XAI gesehen werden kann. Durch Empfehlungen unterstützt es den Nutzer erfolgreich bei der Auswahl geeigneter XAI-Methoden und motiviert ihn zu ihrer tatsächlichen Anwendung.

# Literaturverzeichnis

- Adadi, A. & Berrada, M. (2018), ‘Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)’, *IEEE Access* **6**, 52138–52160.
- AlgorithmWatch (2019), ‘AI Ethics Guidelines Global Inventory’, [Online], Verfügbar unter <https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/>. (Zugriff am: 23.04.2021).
- Angwin, J., Larson, J., Kirchner, L. & Mattu, S. (2016), ‘Machine Bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.’, [Online], Verfügbar unter <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. (Zugriff am: 19.10.2020).
- Antidiskriminierungsstelle des Bundes (2006), ‘Allgemeines Gleichbehandlungsgesetz: AGG’.
- Arrieta, A. B., Rodríguez, N. D., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. & Herrera, F. (2020), ‘Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI’, *Information Fusion* **58**, 82–115.
- Baak, M., Koopman, R., Snoek, H. & Klous, S. (2020), ‘A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics’, *Computational Statistics & Data Analysis* **152**, 107043.
- Bai, Y. & Roth, Z. S. (2019), *Classical and Modern Controls with Microcontrollers*, Advances in Industrial Control, Springer.
- Belle, V. & Papantonis, I. (2020), ‘Principles and Practice of Explainable Machine Learning’, *arXiv preprint arXiv:2009.11698*.
- Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. & Eckersley, P. (2020), Explainable machine learning in deployment, in ‘Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency’, pp. 648–657.
- Boehm, B. W., Brown, J. R. & Lipow, M. (1976), ‘Quantitative Evaluation of Software Quality’, *ICSE ’76: Proceedings of the 2nd international conference on Software engineering* pp. 592–605.
- Böhme, G. (1993), *Fuzzy-Logik: Einführung in die algebraischen und logischen Grundlagen*, Fuzzy-Logik, Springer Berlin Heidelberg, Berlin, Heidelberg.
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. (2016), ‘Man is to computer programmer as woman is to homemaker? debiasing word embeddings’, *Advances in neural information processing systems* **29**, 4349–4357.

- Braun, R., Benedict, M., Wendler, H. & Esswein, W. (2015), Proposal for Requirements Driven Design Science Research, in B. Donnellan, M. Helfert, J. Kenneally, D. VanderMeer, M. Rothenberger & R. Winter, eds, 'New Horizons in Design Science: Broadening the Research Agenda', Vol. 9073, Springer, Cham, pp. 135–151.
- Buber, R. & Holzmüller, H. H. (2009), *Qualitative Marktforschung: Konzepte - Methoden - Analysen*, Gabler-Lehrbuch, 2. Aufl., Gabler, Wiesbaden.
- Carvalho, D. V., Pereira, E. M. & Cardoso, J. S. (2019), 'Machine Learning Interpretability: A Survey on Methods and Metrics', *Electronics* **8**(8), 832.
- Cho, H.-C., Lee, D., Ju, H., Park, H.-C., Kim, H.-Y. & Kim, K. (2017), 'Fire damage assessment of reinforced concrete structures using fuzzy theory', *Applied Sciences* **7**, 518.
- Coca, A. (2020), 'GitHub Issue Comment', [Online], Verfügbar unter <https://github.com/SeldonIO/alibi/issues/221#issuecomment-628049558>. (Zugriff am: 26.01.2021).
- Dastin, J. (2018), 'Amazon scraps secret AI recruiting tool that showed bias against women', [Online], Verfügbar unter <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>. (Zugriff am: 14.10.2020).
- DataRobot (2020), 'Model Interpretability', [Online], Verfügbar unter <https://www.datarobot.com/wiki/interpretability/>. (Zugriff am: 12.10.2020).
- Django Software Foundation (2021), 'Django', Version 3.2.4, Verfügbar unter <https://www.djangoproject.com>.
- Döbel, I., Leis, M., Molina Vogelsang, M., Neustroev, D., Petzka, H., Riemer, A., Rüping, S., Voss, A., Wegele, M. & Welz, J. (2018), 'Maschinelles Lernen: Eine Analyse zu Kompetenzen, Forschung und Anwendung'.
- Doshi-Velez, F. & Kim, B. (2017), 'Towards A Rigorous Science of Interpretable Machine Learning', *arXiv preprint arXiv:1702.08608*.
- Dua, D. & Graff, C. (2017), 'UCI Machine Learning Repository', [Online], Verfügbar unter <https://archive.ics.uci.edu/ml/datasets/adult>. (Zugriff am: 19.10.2020).
- Europäisches Parlament und Rat der Europäischen Union (2016), 'Verordnung (EU) 2016/679: DSGVO'.
- Evans, J. D. (1996), *Straightforward statistics for the behavioral sciences*, Thomson Brooks/Cole Publishing Co.
- Facebook Inc. (2020), 'React', Version 16.14.0, Verfügbar unter <https://reactjs.org>.
- Fahrmeir, L. (2004), *Statistik: Der Weg zur Datenanalyse*, Springer-Lehrbuch, 5. Aufl., Springer, Berlin.



- Federal Office for Information Security (2021), ‘AI Cloud Service Compliance Criteria Catalogue (AIC4)’.
- Formium Inc. (2020), ‘Formik’, Version 2.2.6, Verfügbar unter <https://formik.org>.
- Garcia, S., Luengo, J., Sáez, J. A., López, V. & Herrera, F. (2013), ‘A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning’, *IEEE Transactions on Knowledge and Data Engineering* **25**(4), 734–750.
- Gatsby Inc. (2021), ‘GatsbyJS’, Version 3.5.0, Verfügbar unter <https://www.gatsbyjs.com>.
- Gilpin, L. H., Testart, C., Fruchter, N. & Adebayo, J. (2019), ‘Explaining explanations to society’, *arXiv preprint arXiv:1901.06560*.
- Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. (2015), ‘Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation’, *Journal of Computational and Graphical Statistics* **24**(1), 44–65.
- Google (2020a), ‘Data preprocessing for machine learning: Options and recommendations: Pre-processing data for machine learning’, [Online], Verfügbar unter [https://cloud.google.com/solutions/machine-learning/data-preprocessing-for-ml-with-tf-transform-pt1#preprocessing\\_data\\_for\\_machine\\_learning](https://cloud.google.com/solutions/machine-learning/data-preprocessing-for-ml-with-tf-transform-pt1#preprocessing_data_for_machine_learning). (Zugriff am: 13.11.2020).
- Google (2020b), ‘Introduction to AI Explanations for AI Platform’, [Online], Verfügbar unter <https://cloud.google.com/ai-platform/prediction/docs/ai-explanations/overview>. (Zugriff am: 12.10.2020).
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D. & Giannotti, F. (2018), ‘A Survey Of Methods For Explaining Black Box Models’, *ACM computing surveys (CSUR)* **51**(5), 1–42.
- H2O.ai (2020), ‘Machine Learning Interpretability: Interpretability in H2O Driverless AI’, [Online], Verfügbar unter <https://www.h2o.ai/products-dai-mli/>. (Zugriff am: 12.10.2020).
- Hall, P. & Gill, N. (2019), *An Introduction to Machine Learning Interpretability*, O’Reilly Media, Incorporated.
- Hamon, R., Junklewitz, H. & Sanchez, I. (2020), ‘Robustness and explainability of artificial intelligence’, *Publications Office of the European Union*.
- Han, J., Kamber, M. & Pei, J. (2012), *Data mining: Concepts and techniques*, 3. Aufl., Elsevier Inc., 225 Wyman Street, Waltham, MA 02451, USA.
- Harris, D. & Harris, S. L. (2010), *Digital design and computer architecture*, Morgan Kaufmann.
- Hevner, A. R., March, S. T., Park, J. & Ram, S. (2004), ‘Design science in information systems research’, *MIS Quarterly* (1), 75–105.
- Hinkle, D. E., Wiersma, W. & Jurs, S. G. (2003), *Applied statistics for the behavioral sciences*, Vol. 663, Houghton Mifflin College Division.

- Holsapple, C. W. & Whinston, A. B. (2013), Expert Systems, *in* S. I. Gass & M. C. Fu, eds, 'Encyclopedia of Operations Research and Management Science', Springer US, Boston, MA, pp. 528–532.
- Iarocci, N. (2019), 'Cerberus', Version 1.3.2, Verfügbar unter <https://docs.python-cerberus.org/en/stable>.
- Kaggle Inc. (2021), 'Titanic - Machine Learning from Disaster', [Online], Verfügbar unter <https://www.kaggle.com/c/titanic/data>. (Zugriff am: 18.10.2020).
- Kangur, A. (2020), 'Explainable AI in practice: How committing to transparency made us deliver better AI products', [Online], Verfügbar unter <https://towardsdatascience.com/explainable-ai-in-practice-6d82b77bf1a7>. (Zugriff am: 28.01.2021).
- Kraus, T., Ganschow, L., Eisenträger, M. & Wischmann, S. (2021), *Erklärbare Künstliche Intelligenz - Anforderungen, Anwendungen, Lösungen*.
- Kubeflow (2020), 'Overview of Kubeflow Pipelines: Understanding the goals and main concepts of Kubeflow Pipelines', [Online], Verfügbar unter <https://www.kubeflow.org/docs/pipelines/overview/pipelines-overview>. (Zugriff am: 19.10.2020).
- Kubeflow (2021), 'Building Python function-based components', [Online], Verfügbar unter <https://www.kubeflow.org/docs/components/pipelines/sdk/python-function-components/>. (Zugriff am: 21.06.2021).
- La Vigne, F. (2018), 'Using jupyter notebooks', [Online], Verfügbar unter <https://docs.microsoft.com/en-us/archive/msdn-magazine/2018/february/artificially-intelligent-using-jupyter-notebooks>. (Zugriff am: 11.05.2021).
- Le, S. Q. & Ho, T. B. (2005), 'An association-based dissimilarity measure for categorical data', *Pattern Recognition Letters* **26**(16), 2549–2557.
- Leslie, D. (2019), 'Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector', *SSRN 3403301*.
- Lundberg, S. M. & Lee, S.-I. (2017), 'A Unified Approach to Interpreting Model Predictions', *arXiv preprint arXiv:1705.07874*.
- Mamdani, E. H. & Assilian, S. (1975), 'An experiment in linguistic synthesis with a fuzzy logic controller', *International journal of man-machine studies* **7**(1), 1–13.
- Mamdani, E. H., Efstathiou, H. J. & Sugiyama, K. (1984), Developments in fuzzy logic control, *in* 'The 23rd IEEE Conference on Decision and Control', pp. 888–893.
- McKinsey Global Publishing (2020), 'Global survey: The State of AI in 2020'.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2019), 'A survey on bias and fairness in machine learning', *arXiv preprint arXiv:1908.09635*.
- Miller, T. (2019), 'Explanation in Artificial Intelligence: Insights from the Social Sciences', *Artificial intelligence* **267**, 1–38.

- Miranda, P., Isaias, P. & Crisóstomo, M. (2011), Evaluation of expert systems: The application of a reference model to the usability parameter, *in* ‘International Conference on Universal Access in Human-Computer Interaction’, Vol. 6765, Springer, pp. 100–109.
- Molnar, C. (2019), *Interpretable Machine Learning*. Verfügbar unter <https://christophm.github.io/interpretable-ml-book/>.
- Nissen, V. (2007), ‘Ausgewählte Grundlagen der Fuzzy Set Theorie’, *Reihe Ilmenauer Beiträge zur Wirtschaftsinformatik, Arbeitsbericht Nr. 2007-03*.
- Partnership on AI (2021), ‘Artificial Intelligence Incident Database’, [Online], Verfügbar unter <https://incidentdatabase.ai>. (Zugriff am: 14.04.2021).
- Prince, A. E. R. & Schwarcz, D. (2020), ‘Proxy discrimination in the age of artificial intelligence and big data’, *Iowa Law Review* **105**, 1257–1318.
- Puppe, F. (1991), *Einführung in Expertensysteme*, 2. Aufl., Springer, Berlin, Heidelberg.
- Python Software Foundation (2020), ‘Python Language Reference’, Version 3.9, Verfügbar unter <https://www.python.org>.
- Ribeiro, M. T., Singh, S. & Guestrin, C. (2016), ‘Why should i trust you? Explaining the predictions of any classifier’, *in* ‘Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining’, pp. 1135–1144.
- Riberio, M. T., Singh, S. & Guestrin, C. (2018), ‘Anchors: High precision model-agnostic explanations’, *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Ronacher, A. (2020), ‘Flask’, Version 1.1.2, Verfügbar unter <https://palletsprojects.com/p/flask/>.
- Spinner, T., Schlegel, U., Schafer, H. & El-Assady, M. (2020), ‘explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning’, *IEEE transactions on visualization and computer graphics* **26**(1), 1064–1074.
- The Data Visualisation Catalogue (2019), ‘Proportional Area Charts’, [Online], Verfügbar unter [https://datavizcatalogue.com/methods/area\\_chart.html](https://datavizcatalogue.com/methods/area_chart.html). (Zugriff am: 08.06.2021).
- Uppington, W. (2020), ‘Introducing Truera’, [Online], Verfügbar unter <https://truera.com/introducing-truera/>. (Zugriff am: 12.10.2020).
- van Leekwijck, W. & Kerre, E. E. (1999), ‘Defuzzification: Criteria and classification’, *Fuzzy Sets and Systems* **108**(2), 159–178.
- van Looveren, A. & Klaise, J. (2019), ‘Interpretable Counterfactual Explanations Guided by Prototypes’, *arXiv preprint arXiv:1907.02584*.
- Vilone, G. & Longo, L. (2020), ‘Explainable Artificial Intelligence: A Systematic Review’, *arXiv preprint arXiv:2006.00093*.

- vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R. & Cleven, A. (2009), ‘Reconstructing the Giant - On the Importance of Rigour in Documenting the Literature Search Process’.
- Warner, J. (2019), ‘Scikit-fuzzy’, Version 0.4.2, Verfügbar unter <https://github.com/scikit-fuzzy/scikit-fuzzy>.
- Zadeh, L. A. (1975), ‘Fuzzy logic and approximate reasoning’, *Synthese* 30 (3) pp. 407–428.
- Ziegler, J. & Loepp, B. (2019), Empfehlungssysteme, in T. Kollmann, ed., ‘Handbuch Digitale Wirtschaft’, Springer Reference Wirtschaft, Springer, Wiesbaden, pp. 1–25.

# A. Anhang

## A.1. Leitfaden des Fragenkatalogs zur Anforderungsanalyse

*Anmerkungen:*

- *Kategorie 1 wurde abhängig von den ML-Kenntnissen der Interviewpartner gestellt oder ausgelassen.*
- *Die Personen waren mit dem groben Thema der Arbeit vertraut. Eine konkrete Vorstellung des Empfehlungssystems und seinen Zielen fand zeitlich nach Stellen der Fragen der Kategorie 2 statt.*
- *Um den Personen eine genauere Vorstellung von dem Empfehlungssystem zu geben und eine frühe Iteration (Iteration 2) der grafischen Benutzeroberfläche zu evaluieren, wurde ihnen die vorläufige Ergebnisdarstellung präsentiert.*

### Kategorie 1: Modell Publishing

1. Wie viele Modelle, die in Produktion gehen, sind Black-Box Modelle?
2. Werden transparente White-Box Modelle für die Aufgaben ebenfalls in Betracht gezogen?
3. Welche Informationen werden berücksichtigt, sodass das Modell ein „Blessing“ erhält?
  - Gibt es besondere Metriken, abgesehen von der Leistung des Modells, auf die geachtet wird?
4. Was wird zur Vermeidung von diskriminierenden Modellen getan?
5. Wann findet die Datenanalyse statt und werden die Ergebnisse nach dem initialen Feature Engineering nochmals validiert?

### Kategorie 2: Erfahrung mit XAI

1. Wie häufig wendest Du XAI-Methoden auf Modelle in der Praxis an?
  - Welche XAI Verfahren hast Du bereits angewendet?
2. Welche XAI Verfahren sind Dir bekannt?
  - Was hindert Dich an der Anwendung dieser?
3. Was sind die aktuellen Probleme bei der Auswahl von XAI-Methoden?
4. Auf welche Faktoren achtest Du bei der Auswahl einer XAI-Methode?

*<Vorstellung des XAI-Empfehlungssystems>*

### Kategorie 3: XAI-Empfehlungssystem

1. Was erhoffst Du Dir von dem Empfehlungssystem?

2. Wie würdest Du das Empfehlungssystem am liebsten verwenden (eigenständig/in einer ML-Pipeline)?
3. Was und in welcher Form steht dem System als Eingabeparameter zur Verfügung?
4. Ist ein Zugriff auf die Daten für die Ermittlung der Eingabeparameter des Systems erfahrungsgemäß möglich (bzgl. Datenschutz)?
5. Welche Anforderungen muss das System erfüllen, um es auch tatsächlich anzuwenden?
6. Was muss das System machen um für dich vertrauenswürdig zu sein?
7. Welche Fragen müssen dafür beantwortet werden, um Dein Vertrauen zu gewinnen?
8. Auf welche Faktoren sollte das System bei der Auswahl einer XAI-Methode achten?
  - Gibt es nutzerspezifische Präferenzen, deren Beachtung Dir wichtig sind?
9. Spielt die Berechnungskomplexität einer XAI-Methode im Kontext der Modellentwicklungsumgebung eine Rolle?
  - Muss auf eine geringe Berechnungszeit/Ressourcenverwendung geachtet werden?

**<Vorstellung der Ergebnisdarstellung der Iteration 2>**

**Kategorie 4: Evaluation des Empfehlungsergebnisses der Iteration**

1. Welche Informationen sollte eine Empfehlung beinhalten?
2. Welche Informationen sind für Dich wichtig, um eine Auswahl treffen zu können?
3. In welcher Form soll das Ergebnis präsentiert werden?
4. Wie muss das Ergebnis aussehen, dass Deine Bereitschaft der tatsächlichen Anwendung der Methode erhöht wird?

**A.2. Implementierungsbewertung der ausgewählten XAI-Methoden**

*Anmerkungen:*

- Die verschiedenen Implementierungen wurden im Januar 2021 evaluiert und final für das Empfehlungssystem ausgewählt.
- Für die Auswahl wurden ausschließlich Python Implementierungen berücksichtigt, die keiner Copyleft-Lizenz unterliegen.
- Die Namen der für den XAIR ausgewählten Implementierungen sind fett gedruckt.
- Eigenschaften einer Implementierung, welche die Entscheidung für eine Auswahl in besonderem Maße beeinflusst haben, sind rot gedruckt.
- Die Bewertung der Community fand anhand der „Stars“ und „Forks“ der gesamten GitHub-Projekte statt. Dabei muss beachtet werden, dass sich diese Metriken nicht nur auf die betrachteten XAI-Implementierungen, sondern ggf. auf die ganze Open-Source Bibliothek des jeweiligen Unternehmens (z.B. Alibi von Seldon oder scikit-learn) beziehen.  
Nicht speziell auf die Implementierung bezogene Bewertungen sind daher kursiv gedruckt.

## A.2. Implementierungsbewertung der ausgewählten XAI-Methoden

Tabelle A.1.: Implementierungsvergleich PDP+ICE

PDP + ICE		
Name	PDPbox	scikit-learn
Projektarchiv	<a href="https://github.com/SauceCat/PDPbox">https://github.com/SauceCat/PDPbox</a>	<a href="https://github.com/scikit-learn/scikit-learn">https://github.com/scikit-learn/scikit-learn</a>
Package Version	0.2.0	0.23.2
Lizenz	MIT	BSD 3-Clause License
Dokumentation	Gut	Gut
Community	Star: 547 Fork: 89	Star: 45700 Fork: 21400
Einfache Anwendung	Ja	Ja
Kategorische Features	Möglich	Nein
Unterstützung OHE	Ja, Ermittlung der Spaltennamen notwendig	Nein
Unterstützung LE	Ja	Nein
Zusätzliche Vorteile	<ul style="list-style-type: none"><li>Interaktions-Plots (Grid-Plot, Contour-Plot)</li><li>Weitere Informations-Plots verfügbar (Verteilung der tatsächlichen Vorhersagen für Feature-Werte, durchschnittliche Vorhersage für jeden Feature-Wert)</li></ul>	3D-Interaktions-Plots zweier Features
Nachteile	Nur <i>Scikit-Learn</i> Modell	Multi-Output/ Multi-Class Klassifikationsmodelle werden nicht unterstützt

Tabelle A.2.: Implementierungsvergleich ALE

ALE			
Name	PyALE	ALEPython	Alibi
Projektarchiv	<a href="https://github.com/DanaJomar/PyALE">https://github.com/DanaJomar/PyALE</a>	<a href="https://github.com/blent-ai/ALEPython">https://github.com/blent-ai/ALEPython</a>	<a href="https://github.com/SeldonIO/alibi">https://github.com/SeldonIO/alibi</a>
Package Version	1.0.0.post1	-	0.5.5
Lizenz	MIT	Apache License Version 2.0	Apache License Version 2.0
Dokumentation	Vergleichsweise gut	Schlechte Dokumentation, Kommentare im Code OK	Gut
Community	Star: 3 Fork: 2	Star: 76 Fork: 29	Star: 981 Fork: 126
Einfache Anwendung	Ja	Lokale Installation notwendig (nicht über pip installierbar)	Ja
Kategorische Features	Möglich	Nein	Nein
Unterstützung OHE	Zusätzliche Vorbereitungsschritte, siehe Dokumentation	Nein	Nein
Unterstützung LE	Keine zusätzlichen Vorbereitungsschritte	Nein	Nein
Zusätzliche Vorteile	-	Berechnung mit Monte-Carlo Simulation	-
Nachteile	-	Fehleranfällig (siehe Projektarchiv „Issues“)	-

## A.2. Implementierungsbewertung der ausgewählten XAI-Methoden

Tabelle A.3.: Implementierungsvergleich SHAP

SHAP		
Name	SHAP	Alibi
Projektarchiv	<a href="https://github.com/slundberg/shap">https://github.com/slundberg/shap</a>	<a href="https://github.com/SeldonIO/alibi">https://github.com/SeldonIO/alibi</a>
Package Version	0.38.1	0.5.5
Lizenz	MIT	Apache License Version 2.0
Dokumentation	Gut	Gut
Community	Star: 12600 Fork: 1900	Star: 981 Fork: 126
Einfache Anwendung	Etwas aufwendiger wenn OHE	Ja
Kategorische Features	Möglich	Möglich
Unterstützung OHE	Durch Aufsummieren oder Gruppenbildung	Angabe der OHE Feature-Indexe pro Feature notwendig
Unterstützung LE	Ja	Ja
Zusätzliche Vorteile	<ul style="list-style-type: none"> <li>Fundiert: Original-Implementierung der Veröffentlichung von Lundberg &amp; Lee (2017)</li> <li>Einzige Implementierung</li> <li>Modellspezifische Varianten mit besserer Leistung verfügbar</li> </ul>	<ul style="list-style-type: none"> <li>Wrapper um Original-Implementierung von Lundberg &amp; Lee (2017), weist daher dieselben Vorteile auf</li> <li>Explizites Gruppieren von OHE Features (spart Berechnungszeit)</li> </ul>
Nachteile	OHE Features werden nicht explizit als solche behandelt	Original SHAP Package muss trotzdem importiert werden

Tabelle A.4.: Implementierungsvergleich Anchors

Anchors		
Name	Anchors	Alibi
Projektarchiv	<a href="https://github.com/marcotcr/anchor">https://github.com/marcotcr/anchor</a>	<a href="https://github.com/SeldonIO/alibi">https://github.com/SeldonIO/alibi</a>
Package Version	0.0.2.0	0.5.5
Lizenz	BSD 2-Clause „Simplified“ License	Apache License Version 2.0
Dokumentation	OK	Gut
Community	Star: 634 Fork: 91	Star: 981 Fork: 126
Einfache Anwendung	Ja	Ja
Kategorische Features	Möglich	Angabe der OHE Feature-Indexe pro Feature notwendig
Unterstützung OHE	Möglich wenn Angabe einer Kodierungsfunktion oder Kodierung innerhalb der Vorhersagefunktion	Möglich wenn Kodierung innerhalb der Vorhersagefunktion
Unterstützung LE	Ja	Ja
Zusätzliche Vorteile	-	-
Nachteile	Keine benutzerdefinierte Diskretisierung	-



Tabelle A.5.: Implementierungsvergleich Counterfactual Explanations

Counterfactual Explanations		
Name	Alibi (Counterfactuals guided by Prototypes (CFProto))	DiCE (Diverse Counterfactual Explanations)
Projektarchiv	<a href="https://github.com/SeldonIO/alibi">https://github.com/SeldonIO/alibi</a>	<a href="https://github.com/interpretml/DiCE">https://github.com/interpretml/DiCE</a>
Package Version	0.5.5	0.4
Lizenz	Apache License Version 2.0	MIT
Dokumentation	Gut	Gut
Community	Star: 981 Fork: 126	Star: 526 Fork: 78
Einfache Anwendung	Ja	Ja
Kategorische Features	Möglich	Möglich
Unterstützung OHE	Möglich wenn Angabe einer Category Map (automatische Generierung möglich)	Nur OHE bei implementierungsspezifischem <i>DiCE Data</i> Objekt
Unterstützung LE	Ja	Ja
Zusätzliche Vorteile	<i>TensorBoard</i> Unterstützung	<ul style="list-style-type: none"> <li>Zusätzliche Angabe vieler Parameter (Anzahl der CFs, Diversität, ...)</li> </ul>
Nachteile	Diskretisierung numerischer Features notwendig	<ul style="list-style-type: none"> <li>Modell muss differenzierbar sein</li> <li>Nur <i>TensorFlow</i> und <i>PyTorch</i> Klassifizierer unterstützt</li> <li>Verwendung des implementierungs- spezifischen <i>DiCE Data</i> Objekts obligatorisch</li> </ul>

Tabelle A.6.: Implementierungsvergleich PFI

PFI			
Name	Scikit-Learn	ELI5	mlxtend
Projektarchiv	<a href="https://github.com/scikit-learn/scikit-learn">https://github.com/scikit-learn/scikit-learn</a>	<a href="https://github.com/eli5-org/eli5">https://github.com/eli5-org/eli5</a>	<a href="https://github.com/rasbt/mlxtend">https://github.com/rasbt/mlxtend</a>
Package Version	0.23.2	0.10.1	0.17.3
Lizenz	BSD 3-Clause License	MIT	New BSD Open Source License
Dokumentation	Gut	OK	OK
Community	Star: 45700 Fork: 21400	Star: 30 Fork: 4	Star: 3500 Fork: 712
Einfache Anwendung	Eigenständige Visualisierung notwendig	Ja	Eigenständige Visualisierung notwendig
Kategorische Features	Nur, wenn Modell Teil einer <i>Scikit-Learn</i> Pipeline (Preprocessing Schritte automatisch ausgeführt)	Möglich	Möglich
Unterstützung OHE	Wenn Modell Teil einer <i>Scikit-Learn</i> Pipeline: Berechnung gebündelte Wichtigkeit aller OHE-Features	Manuelle Addition einzelner OHE Feature-Wichtigkeiten notwendig	Manuelle Addition einzelner OHE Feature-Wichtigkeiten notwendig
Unterstützung LE	Siehe "Kategorische Features"	Ja	Ja
Zusätzliche Vorteile	<ul style="list-style-type: none"> <li>Große Auswahl an Leistungsmetriken zur Berechnung der Feature-Wichtigkeit</li> </ul>	<ul style="list-style-type: none"> <li>Große Auswahl an Leistungsmetriken zur Berechnung der Feature-Wichtigkeit</li> <li>Ansprechende Visualisierung</li> </ul>	<ul style="list-style-type: none"> <li>Verwendung benutzerdefinierter Leistungsmetriken möglich</li> </ul>
Nachteile	<ul style="list-style-type: none"> <li>Eigenständige Visualisierung notwendig</li> </ul>	<ul style="list-style-type: none"> <li>Bei OHE: Visualisierungsfunktion stellt nur einzelne OHE Features dar (nicht gebündelte)</li> </ul>	<ul style="list-style-type: none"> <li>Eigenständige Visualisierung notwendig</li> <li>Vergleichsweise wenige Rückgaben</li> </ul>

## A.3. Formulierungen bei der Eingabe benötigter Parameter

### A.3.1. Eingabeformulierung der Ausschlusskriterien

Die Bewertungen der XAI-Methoden wurden so formuliert, dass ein Methodenausschluss binär abgefragt werden kann. Die Anwendung einer Methode ist demnach möglich, wenn die Eingabevariable größer als oder gleichgroß wie die Kriteriumsbewertung ist. Andernfalls wird die Methode aus der Empfehlungsergebnismenge exkludiert.

Tabelle A.7.: Formulierungen der Erfragung der Ausschlusskriterien

<b>Ausschlusskriterium</b>	<b>Eingabeparameterbewertung</b>	<b>Methodenbewertung</b>
<b>Verfügbarkeit des Modells</b>	Ist der Zugriff auf das Modell gegeben (oder nur auf die Vorhersagefunktion)?	Ist der Zugriff auf das Modell für die Anwendung des Verfahrens notwendig (oder reicht der Zugriff auf die Vorhersagefunktion aus)?
<b>Klassifikationsaufgabe</b>	Liegt eine Klassifikationsaufgabe (oder eine Regressionsaufgabe) vor?	Ist die Methode auf die Anwendung bei Klassifikationsaufgaben beschränkt (oder ist sie bei Regressionsaufgaben möglich)?
<b>Erhalt der Klassenwahrscheinlichkeiten</b>	Bei einer Klassifikationsaufgabe: Gibt das Modell die Klassenwahrscheinlichkeiten zurück?	Ist der Erhalt der Klassenwahrscheinlichkeiten für die Anwendung der Methode notwendig?
<b>Zugriff auf Labels</b>	Sind die Trainingslabels verfügbar?	Ist der Zugriff auf die Trainingslabels für die Anwendung der Methode notwendig?
<b>Zugriff auf Preprocessing Operationen</b>	Sind die Preprocessing Operationen verfügbar und reversibel?	Ist eine inverse Transformation der Ausgabedaten (Rückgängigmachen der Skalierung/Kodierung durch Vorhandensein der Preprocessing-Operationen) für die Anwendung der Methode notwendig?

## A.3.2. Eingabeformulierung der eignungsbeeinflussenden Kriterien

Tabelle A.8.: Formulierungen der Erfragung eignungsbeeinflussender Eingabeparameter

Kriterium	Eingabeparameterbewertung	Methodenbewertung
<b>Präferenz einer globalen Erklärung</b>	Werden XAI-Methoden bevorzugt, die globale Erklärungen des gesamten Systems, unabhängig von einer spezifischen Eingabe, liefern?	Liefert die Methode eine globale Erklärung?
<b>Präferenz einer lokalen Erklärung</b>	Werden XAI-Methoden bevorzugt, die lokale Erklärungen liefern, d.h. die die Vorhersage einer spezifischen Dateninstanz erklären?	Liefert die Methode eine lokale Erklärung?
<b>FOI</b>	Welche Features bieten voraussichtlich das Potential zur Diskriminierung und benötigen daher besondere Aufmerksamkeit (Beispiele: Geschlecht, Rasse)?	(*)
<b>Korrelation</b>	Wie stark ist die Korrelation des Datensatzes (unter Berücksichtigung der Anzahl korrelierender Features und der Stärke der Korrelationen)?	Wie geht die Methode mit Korrelationen der Eingabefeatures um?
<b>Korrelation der FOI</b>	Wie stark ist die Korrelation der interessanten Features (unter Berücksichtigung der Anzahl und Stärke der Korrelationen mit allen Eingabe-Features)	Siehe „Korrelation“
<b>Diskretisierbarkeit</b>	Wie gut diskretisierbar sind die numerischen Features?	Wie geht die Methode mit schlecht diskretisierbaren Features um?
<b>Diskretisierbarkeit der FOI</b>	Wie gut diskretisierbar sind die numerischen FOI?	Siehe „Diskretisierbarkeit“
<b>Präferenzen der Performance</b>	Wie aufwendig darf die Berechnung der XAI-Methode sein (Berechnungskomplexität bezüglich Ressourcenverbrauch)?	(*)

Forsetzung - Formulierungen der Erfragung eignungsbeeinflussender Eingabeparameter		
Kriterium	Eingabeparameterbewertung	Methodenbewertung
Anzahl der Features	Wie viele Features hat der Datensatz (inklusive One-Hot kodierte)?	Welche Auswirkung hat die Anzahl der Features auf die Berechnungskomplexität der Methode?
Zugriffszeit Modell/ Vorhersagefunktion	Wie lange dauert der Zugriff auf das Modell bzw. die Vorhersagefunktion?	Wie oft greift die Methode während einer Anwendung auf das Modell bzw. die Vorhersagefunktion zu?
Vorbereitungsaufwand	Wie viel Zeit steht zur Methodenvorbereitung und Einarbeitung zur Verfügung?	Wie lange dauert die Vorbereitung bzw. die Einarbeitungszeit für die Methodenanwendung?
Vorhandensein ordinaler Features	Enthält der Datensatz ordinale Features?	(*)

(\*) Eigenständig nicht relevant für Bewertung der Methode, allerdings relevant für oder in Kombination mit anderen Eingabewerten.

## A.4. Empfehlung zur Ausführung der Preprocessing Operationen

Man kann sagen, dass Vorverarbeitungsschritte in non-destruktive Operationen, bei denen die Repräsentation des Features vor der Transformation wiederherstellbar ist, und destruktive Operationen, nicht bijektiv und daher nicht rückgängig machbar sind, kategorisiert werden können.

Non-destruktive Verarbeitungsschritte sind bspw.:

- Änderung des Datentyps ohne Informationsverlust(String zu Kategorie/Zahl), Integer zu Float)
- Datenbereinigung in Form von Entfernen von Dateninstanzen oder Features ohne Informationsgehalt, mit vielen korrupten, fehlenden oder invaliden Werten
- Konstruieren neuer Features aus bestehenden

Destruktive Verarbeitungsschritte sind bspw.:

- Änderung des Datentyps mit Informationsverlust(Float zu Integer)
- Aggregation von Werten zu Buckets (numerisch) oder Kategorie („hellblau“und „marineblau“zu „blau“)
- Ersetzen fehlender Werte
- Ersetzen von Feature(s) durch konstruierte(s) Feature(s)

Es ist empfehlenswert, zuerst gebündelt alle non-destruktiven und anschließend die destruktiven Operationen durchzuführen. Diese Aufteilung des Preprocessings in diese Kategorien ermöglicht ein Vergleich der nicht verfälschten aber strukturierten, vorbereiteten Daten (Ergebnis non-destruktiver Operationen) mit den Engineered Features, die dem Modell als Input dienen. So kann nachvollzogen werden, welche feature-verändernden Schritte vorgenommen wurden und sichergestellt werden, dass die Struktur der Daten nicht grundlegend verändert oder verfälscht wird, dass bspw. starke Abhängigkeiten nicht gebrochen werden. Eventuelle Fehler, die sich von den Daten auch auf das daraus resultierende Modell übertragen, können aufgedeckt werden.

Anmerkung: Obwohl die Standardisierung, Normalisierung durch Speichern und inverse Anwendung des Skalierers allerdings auch non-destruktiv ist, sollte sie zugunsten der Verständlichkeit der Daten in einer Datenanalyse erst am Ende angewendet werden.

## A.5. Vergleich der Korrelationskoeffizienten $\Phi_K$ und Pearson's $\rho$

Es wurde ein Vergleich der bivariaten Korrelationsberechnung von  $\Phi_K$  mit dem Pearson Korrelationskoeffizienten  $\rho$  durchgeführt, um die Eignung von  $\Phi_K$  bei schlecht diskretisierbaren Features beurteilen zu können.

Dafür wurden künstlich Feature Verteilungen mit 1000 bzw. 10000 Dateninstanzen erzeugt:

- drei zufällig erzeugte, aber korrelierende Feature Verteilungen (*corr0*, *corr1*, *corr2*)
- zwei Normalverteilungen (*normal1*, *normal2*)
- zwei rechtssteile Beta-Verteilungen (*beta1*, *beta2*) und eine etwas linkssteile (*beta3*)
- eine sehr schlecht diskretisierbare, linkssteile Pareto II/Lomax Verteilung (*pareto1*)

Der Unterschied der mit den beiden Verfahren berechneten Korrelationswerte wurde in mehreren Iterationen mit jeweils neu generierten Verteilungen gemessen. In Tabelle A.9 sind beispielhaft die deskriptiven Statistiken der Verteilungen von einer Iteration mit 1000 Dateninstanzen aufgeführt.

Tabelle A.9.: Deskriptive Statistik der Verteilungen einer Iteration

Statistik	Verteilung								
	corr0	corr1	corr2	normal1	normal2	beta1	beta2	beta3	pareto1
Anzahl	1000	1000	1000	1000	1000	1000	1000	1000	1000
Mittelwert	-0.025041	-0.046004	0.014751	-0.002431	70.002601	0.832694	0.979953	0.019512	13.824364
Standard- verteilung	1.796044	2.305233	1.089774	0.098509	0.206954	0.146951	0.020575	0.006151	15.557355
Minimum	-5.071324	-7.926681	-3.104210	-0.316295	69.415453	0.068021	0.838459	0.005306	7.003134
0.25-Quantil	-1.289676	-1.596838	-0.752957	-0.066424	69.859724	0.761599	0.972434	0.015026	8.194093
0.5-Quantil	-0.084449	-0.020072	0.046709	-0.002673	69.992850	0.875142	0.985890	0.018975	9.993595
0.75-Quantil	1.202508	1.445011	0.795691	0.061328	70.136765	0.948198	0.994096	0.023100	13.914955
Maximum	5.246396	6.381660	3.152945	0.357908	70.598018	0.999752	0.999985	0.046878	331.668607

Die durchschnittliche Differenz bei Verteilungen von 1000 Dateninstanzen über 50 Iterationen hinweg, ist in Abbildung A.1 zu sehen. Der maximale Unterschied der durchschnittlichen Abweichung von  $\Phi_K$  und  $\rho$  beträgt dabei 0.069357.

Der gleiche Test wurde mit Verteilungen mit 10000 Instanzen über 500 Iterationen durchgeführt und ist in Abbildung A.2 zu sehen. Die maximale Differenz ist dabei 0.028507.

Der Unterschied zwischen  $\Phi_K$  und  $\rho$  bei schlecht diskretisierbaren Verteilungen ist somit marginal.

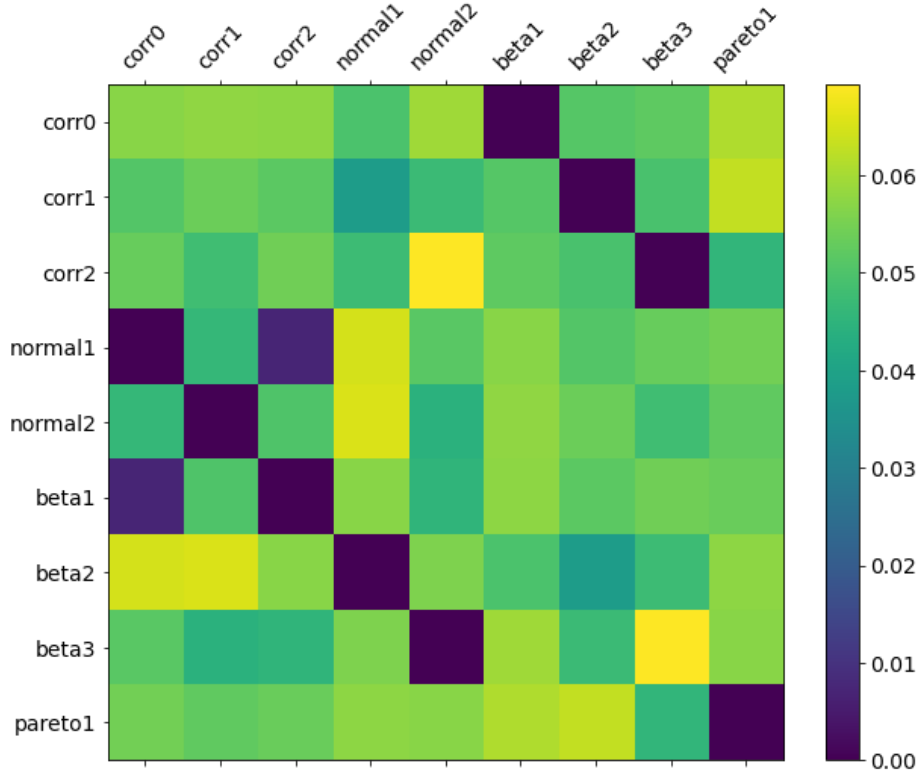


Abbildung A.1.: Durchschnittliche Differenz der Korrelationsstärken von  $\Phi_K$  und Pearson's  $\rho$  über 50 Iterationen zwischen Verteilungen mit jeweils 1000 Dateninstanzen

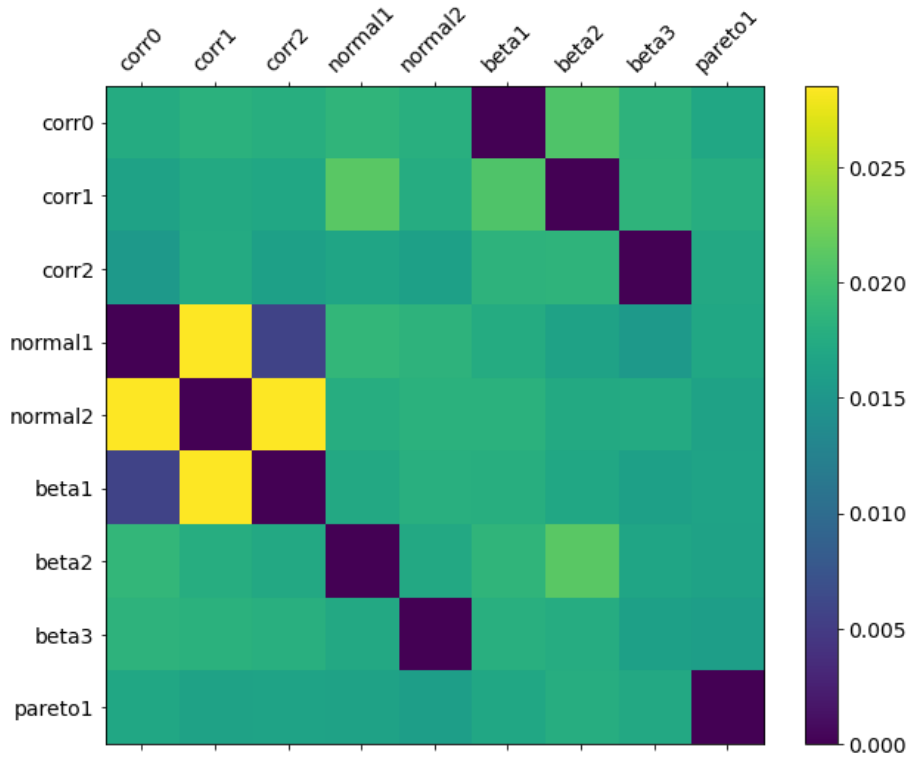


Abbildung A.2.: Durchschnittliche Differenz der Korrelationsstärken von  $\Phi_K$  und Pearson's  $\rho$  über 500 Iterationen zwischen Verteilungen mit jeweils 10000 Dateninstanzen

## A.6. Beispiele der automatisierten Datenanalyse

Nachfolgend werden einige Beispiele zur (automatisierten) Berechnung der datenbezogenen Systemeingaben des XAIRs vorgestellt.

Der Code sowie weitere Beispiele sind in dem Jupyter Notebook *data\_analysis\_concept.ipynb* unter <https://github.com/viadee/xair> zu finden.

### A.6.1. Berechnung der Korrelation

Die Ergebnisse des Konzepts der automatisierten Korrelationsberechnung werden anhand des UCI Adult Datensatzes (Dua & Graff 2017), des Titanic Datensatzes (Kaggle Inc. 2021) und eines konstruierten Beispiels vorgestellt. Dies soll im Speziellen die Halbierung des MADs bei der Aggregation der feature-spezifischen, globalen Korrelationswerte rechtfertigen.

Für jedes Beispiel werden zunächst die gerundeten globalen Korrelationswerte pro Feature (Ergebnis des  $g_K$ ) aufgeführt. Anschließend werden diese den Zugehörigkeitsfunktionen der Korrelation (Tabelle 4.5, S. 52) zugeordnet und die Summe der Zugehörigkeitswerte der Feature-Korrelationen pro Zugehörigkeitsfunktion visualisiert. Dies gibt Aufschluss über die ungefähre Korrelation des Gesamtdatensatzes.

Ihre anschließende Berechnung erfolgt anhand einer Addition des Durchschnitts mit dem halbierten MAD, siehe Gleichung 11 (S. 47:  $corr = \bar{x} + \frac{1}{2}MAD$ ).

Zum Vergleich wird die resultierende Korrelationseinschätzung des Durchschnitts und des unveränderten MADs aufgeführt ( $corr = \bar{x} + MAD$ ).

Wie vor allem anhand des dritten Beispiels ersichtlich, liefert die Addition mit dem unveränderten MAD eine zu hohe fuzzy Einschätzung der Gesamtkorrelation. Obwohl 6 der 13 Features eine Korrelation  $\leq 0.1$  aufweisen, was laut den Korrelationsinterpretationen in Tabelle 4.2 (S. 46) „sehr schwach“ (VL) ist, wird der Datensatz als „stark“ (H) bis „sehr stark“ (VH) korrelierend eingeschätzt.

Auch bei anderen Datensätzen ist die Skalierung durch eine MAD-Halbierung angebracht und spiegelt eine menschliche Einschätzung der Gesamtkorrelation des Datensatzes besser wider.



**Beispiel: UCI Adult Datensatz****Eingabe**

Gerundete globale Korrelationswerte pro Feature (Ergebnis des  $g_K$ ), siehe Abbildung 4.5 (S. 45):

[ 0.974, 0.962, 0.920, 0.879, 0.853, 0.756,  
0.690, 0.639, 0.358, 0.232, 0.223, 0.145 ]

**Eigenschaften der Eingaben**

Zuordnung der globalen Feature-Korrelationen zu den Zugehörigkeitsfunktionen der Korrelation (Tabelle 4.5, S. 52), visualisiert in Abbildung A.3:

Fuzzy-Menge	Anzahl der Features pro Zugehörigkeitsfunktion	Summe der Zugehörigkeitswerte aller Features pro Zugehörigkeitsfunktion
<b>VL</b>	3	2.453
<b>L</b>	3	1.547
<b>M</b>	2	0.705
<b>H</b>	5	2.982
<b>VH</b>	5	4.313

Tabelle A.10.: Zuordnung der  $g_K$  Werte des UCI Adult Datensatzes zu den Zugehörigkeitsmengen der Korrelation

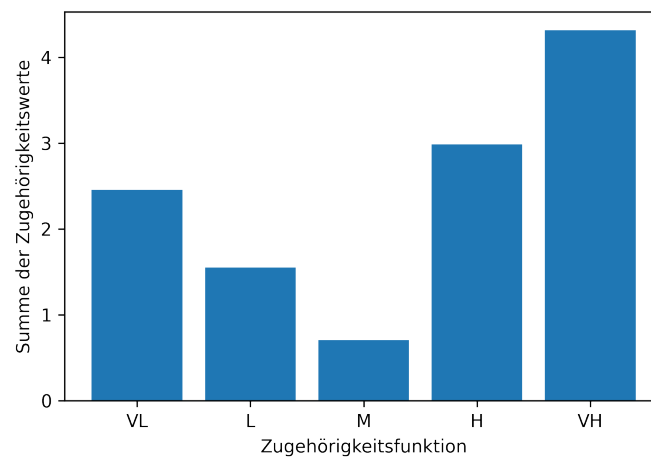


Abbildung A.3.: Summe der Zugehörigkeitswerte der Korrelationen aller Features des UCI Adult Datensatzes pro Zugehörigkeitsfunktion

**Berechnung**

<b>Durchschnitt (<math>\bar{x}</math>):</b>	0.636	$\bar{x} + \frac{1}{2}MAD$	<b>Crisp</b>	0.768
<b>MAD:</b>	0.264		<b>Fuzzy</b>	H
		$\bar{x} + MAD$	<b>Crisp</b>	0.9
			<b>Fuzzy</b>	VH

**Beispiel: Titanic Datensatz****Eingabe**

Gerundete globale Korrelationswerte pro Feature (Ergebnis des  $g_K$ ), siehe Abbildung 4.5:

[ 0.786, 0.383, 0.565, 0.706, 0.607, 0.698, 0.633]

**Eigenschaften der Eingaben**

Zuordnung der globalen Feature-Korrelationen zu den Zugehörigkeitsfunktionen der Korrelation (Tabelle 4.5, S. 52), visualisiert in Abbildung A.4:

Fuzzy-Menge	Anzahl der Features pro Zugehörigkeitsfunktion	Summe der Zugehörigkeitswerte aller Features pro Zugehörigkeitsfunktion
<b>VL</b>	0	0.000
<b>L</b>	1	1.000
<b>M</b>	4	2.624
<b>H</b>	5	3.376
<b>VH</b>	0	0.000

Tabelle A.11.: Zuordnung der  $g_K$  Werte des Titanic Datensatzes zu den Zugehörigkeitsmengen der Korrelation

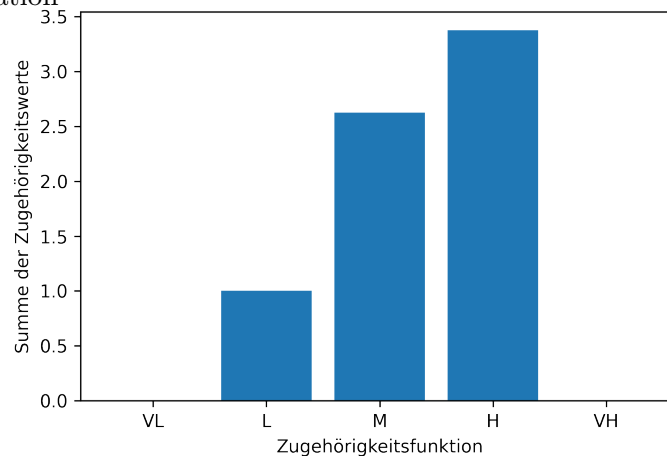


Abbildung A.4.: Summe der Zugehörigkeitswerte der Korrelationen aller Features des Titanic Datensatzes pro Zugehörigkeitsfunktion

**Berechnung**

<b>Durchschnitt (<math>\bar{x}</math>):</b>	0.625	$\bar{x} + \frac{1}{2}MAD$	<b>Crisp</b>	0.671
<b>MAD:</b>	0.092		<b>Fuzzy</b>	M - H
		$\bar{x} + MAD$	<b>Crisp</b>	0.717
			<b>Fuzzy</b>	H

## Konstruiertes Beispiel

### Eingabe

Willkürliche globale Korrelationswerte pro Feature (Ergebnis des  $g_K$ ):

[ 0.10, 0.10, 0.10, 0.10, 0.10, 0.00, 0.90, 0.90, 0.90, 0.85, 0.80, 0.80 ]

### Eigenschaften der Eingaben

Zuordnung der globalen Feature-Korrelationen zu den Zugehörigkeitsfunktionen der Korrelation (Tabelle 4.5, S. 52), visualisiert in Abbildung A.5:

Fuzzy-Menge	Anzahl der Features pro Zugehörigkeitsfunktion	Summe der Zugehörigkeitswerte aller Features pro Zugehörigkeitsfunktion
<b>VL</b>	6	6.0
<b>L</b>	0	0.0
<b>M</b>	0	0.0
<b>H</b>	3	2.5
<b>VH</b>	4	3.5

Tabelle A.12.: Zuordnung beispielhafter globaler Korrelationswerte zu den Zugehörigkeitsmengen der Korrelation

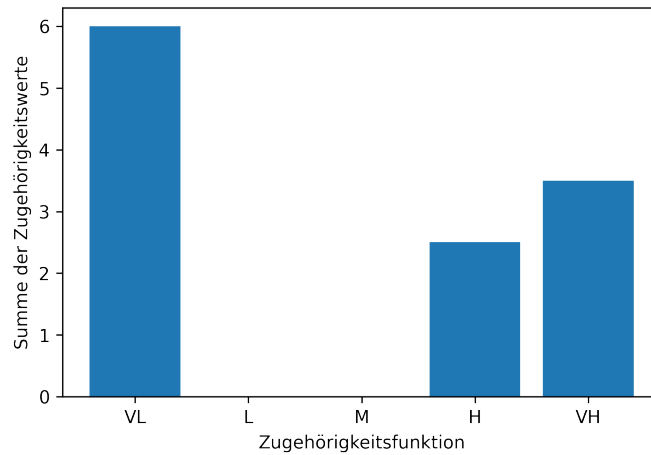


Abbildung A.5.: Summe der Zugehörigkeitswerte der Korrelationen aller Features des konstruierten Datensatzes pro Zugehörigkeitsfunktion

### Berechnung

<b>Durchschnitt (<math>\bar{x}</math>):</b>	0.471	$\bar{x} + \frac{1}{2}MAD$	<b>Crisp</b>	0.665
<b>MAD:</b>	0.388		<b>Fuzzy</b>	M - H
		$\bar{x} + MAD$	<b>Crisp</b>	0.858
			<b>Fuzzy</b>	H - VH

### A.6.2. Berechnung der Diskretisierbarkeit

Zunächst werden die Ergebnisse der automatisierten Berechnung der Diskretisierbarkeit auf Feature-Ebene vorgestellt. Anschließend werden die globalen Diskretisierbarkeitswerte einiger Beispiele genannt.

#### Diskretisierbarkeit eines Features

Die Berechnung der Diskretisierbarkeit wird anhand der in Abbildung 4.6 (Seite 48) aufgeführten Features „Capital Gain“ und „Age“ des UCI Adult Datensatzes präsentiert. Außerdem findet die Berechnung zur Veranschaulichung mit dem Feature „Hours per Week“ desselben Datensatzes statt, dessen Verteilung in dem Histogramm in Abbildung A.6 zu sehen ist.

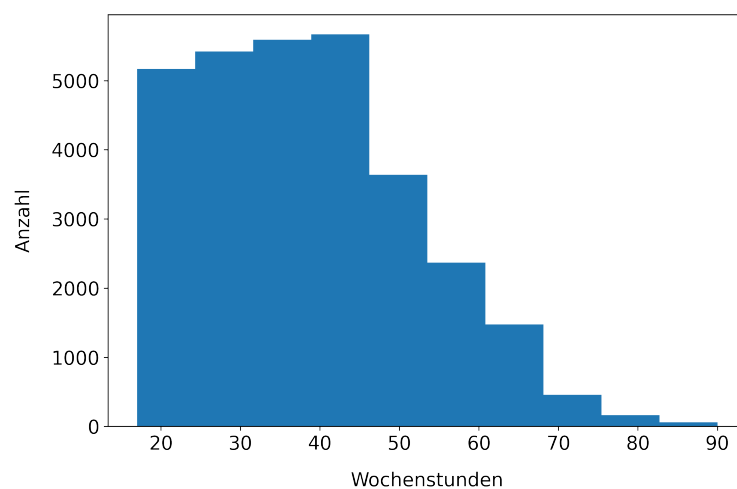


Abbildung A.6.: EqualWidth Binning des Features „Hours per Week“ des UCI Adult Datensatzes

In Tabelle A.13 sind die für die Berechnung relevanten Charakteristika aufgeführt. Wie ersichtlich besitzt das sehr schlecht diskretisierbare Feature „Capital Gain“ einige leere Bins. Auch das Feature „Hours per Week“ gilt als schlecht diskretisierbar, da es identische Bin-Grenzen aufweist. Das ist der Tatsache geschuldet, dass die Anzahl der Dateninstanzen, die eine Arbeitszeit von etwa 40 Stunden pro Woche aufweisen, sehr hoch ist.

#### Diskretisierbarkeit eines Datensatzes

Mit der in Kapitel 4.3.2 erläuterten und prototypisch im Jupyter Notebook implementierten Analyse der Diskretisierbarkeit ergibt sich für den UCI Adult Datensatz eine Diskretisierbarkeit von -1.446, was „schlecht“ ist ( $\mu_L(discr) = 1.0$ ). Dabei wurden alle Features als numerisch betrachtet. Dieses Ergebnis ist sinnvoll, da 11 der 12 Features, alle außer „Age“, Bin-Grenzen mit denselben Werten aufweisen. Dies ist entweder leeren Bins geschuldet, oder bspw. im Falle der Arbeitsstunden pro Woche, „zu vollen“.

Für die Aggregation der einzelnen featurespezifischen Werte der Diskretisierbarkeit wird erneut eine Kombination des Durchschnitts mit dem halbierten MAD verwendet.

	Capital Gain	Age	Hours per Week
Bin-Grenzen	[ 0, 0, 0, 0, 0, 0, 0, 0, 0, 99999 ]	[ 17, 22, 26, 30, 33, 37, 41, 45, 50, 58, 90 ]	[ 1, 24, 35, 40, 40, 40, 40, 40, 48, 55, 99 ]
Bin-Breiten	[ 0, 0, 0, 0, 0, 0, 0, 0, 0, 99999 ]	[ 5, 4, 4, 4, 3, 4, 4, 4, 5, 8, 32 ]	[ 23, 11, 5, 0, 0, 0, 0, 0, 8, 7, 44 ]
Prozentuale Bin-Breite	[ 0, 0, 0, 0, 0, 0, 0, 0, 0, 1 ]	[ 0.068, 0.055, 0.055, 0.041, 0.055, 0.055, 0.055, 0.068, 0.110, 0.438 ]	[ 0.235, 0.112, 0.051, 0.000, 0.000, 0.000, 0.000, 0.082, 0.071, 0.449 ]
MAD der prozentualen Bin-Breiten	0.18	0.07	0.099
Diskretisierbarkeit ( <i>discr</i> )	-8.0	3.0	0.1
Fuzzy Diskretisierbarkeit	$\mu_L(discr) = 1.0$	$\mu_L(discr) = 0.4$ $\mu_M(discr) = 0.6$	$\mu_L(discr) = 0.98$ $\mu_M(discr) = 0.02$

Tabelle A.13.: Die Diskretisierbarkeit betreffende Charakteristika der Features „Capital Gain“, „Age“ und „Hours per Week“

Ihre Sinnhaftigkeit der Addition wurde durch weitere Tests mit diversen MADs der prozentualen Bin-Breiten evaluiert. Nachfolgend werden zwei Beispiele aufgeführt:

#### Beispiel 1: Mittelgut diskretisierbares Feature

MAD pro Feature	[ 0.01, 0.01, 0.01, 0.01, 0.01, 0.12, 0.12, 0.12, 0.12, 0.12, 0.12, 0.12, 0.12 ]
Eingabe in XAIR (Skaliert auf Diskursuniversum)	[9, 9, 9, 9, 9, 9, -2, -2, -2, -2, -2, -2, -2]
Durchschnitt ( $\bar{x}$ )	2.231
MAD	5.207
Diskretisierbarkeit ( $\bar{x} + \frac{1}{2}MAD$ )	4.834
Fuzzy Diskretisierbarkeit	$\mu_L(discr) = 0.033$ $\mu_M(discr) = 0.967$

#### Beispiel 2: Gut diskretisierbares Feature

MAD pro Feature	[ 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.05, 0.05, 0.05, 0.05 ]
Eingabe in XAIR (Skaliert auf Diskursuniversum)	[9, 9, 9, 9, 9, 9, 9, 9, 5, 5, 5, 5]
Durchschnitt ( $\bar{x}$ )	7.545
MAD	1.851
Diskretisierbarkeit ( $\bar{x} + \frac{1}{2}MAD$ )	8.471
Fuzzy Diskretisierbarkeit	$\mu_M(discr) = 0.306$ $\mu_H(discr) = 0.868$

## A.7. Übersicht der Fuzzy-Regeln

Die Namen der Eingabevariablen, wie sie in der Implementierung des XAIRs verwendet werden, lauten:

- **Initialisierungsregel:** `init`
- **Vorhandensein FOI:** `foi_available`
- **Korrelation:** `corr`
- **Korrelation der FOI:** `corr_foi`
- **Diskretisierbarkeit:** `discr`
- **Diskretisierbarkeit der FOI:** `discr_foi`
- **Anzahl der Features:** `num_feat`
- **Präferenzen der Leistung:** `perf_pref`
- **Zugriffszeit des Modells/der Vorhersagefunktion:** `dur_call`
- **Vorhandensein ordinaler Features:** `ordinal_feat`
- **Vorbereitungsaufwand:** `prep_time`
- **Präferenz einer globalen Erklärung:** `scope_global`
- **Präferenz einer lokalen Erklärung:** `scope_local`

Die Fuzzy-Variablen werden dabei innerhalb des *scikit-fuzzy* Systems folgendermaßen dargestellt:

```
<Fuzzy-Variable>[<Term>]@<Gewichtung>\%
```

Dabei sind nur reduzierte Gewichtungen der Ausgabevariablen explizit angegeben.

Die Liste der 38 Fuzzy-Regeln, die nicht ausschließlich die Ausgabevariable *NOOP* sondern XAI-Methoden bewerten, ist nachfolgend aufgeführt:

```
init[True]    --> [PDP + ICE[M], ALE[M], PFI[M], SHAP[M], Anchors[M], CFProto[M]]
corr[L]       --> [PDP + ICE[VH], ALE[L]]
corr[M]       --> [PFI[L], SHAP[L], Anchors[L]@0.60%, CFProto[L]@0.60%]
corr[H]       --> [PDP + ICE[L], ALE[VH], PFI[VL], SHAP[VL], Anchors[L], CFProto[L]]
discr[L]      --> [ALE[L]@0.60%, Anchors[VL]]
discr[M]      --> [Anchors[L]]
corr[L] AND discr[L]
               --> [ALE[VL]@0.70%]
corr[L] AND discr[H]
               --> [ALE[L]@0.50%]
corr[M] AND discr[H]
               --> [ALE[H]@0.70%]
corr[H] AND discr[M]
               --> [ALE[H]@0.70%]
corr[H] AND discr[H]
               --> [ALE[VH]@0.70%]
foi_available[True] AND corr_foi[L]
               --> [PDP + ICE[VH], ALE[L]]
```

```

foi_available[True] AND corr_foi[M]
    --> [PDP + ICE[L], ALE[H]]
foi_available[True] AND corr_foi[H]
    --> [PDP + ICE[VL], ALE[VH], PFI[L], SHAP[L], Anchors[L]@0.60%,
        CFProto[L]@0.60%]
foi_available[True] AND discr_foi[L]
    --> [ALE[L], Anchors[VL]]
foi_available[True] AND discr_foi[M]
    --> [Anchors[L]]
foi_available[True] AND corr_foi[L] AND discr_foi[L]
    --> [ALE[L]@0.60%]
foi_available[True] AND corr_foi[L] AND discr_foi[H]
    --> [ALE[L]@0.60%]
foi_available[True] AND corr_foi[M] AND discr_foi[M]
    --> [ALE[L]@0.60%]
foi_available[True] AND corr_foi[L] AND discr_foi[L]
    --> [ALE[VL]]
foi_available[True] AND corr_foi[H] AND discr_foi[M]
    --> [ALE[H]]
foi_available[True] AND corr_foi[H] AND discr_foi[H]
    --> [ALE[VH]]
perf_pref[M] AND dur_call[M]
    --> [Anchors[L]@0.50%, CFProto[VL]@0.50%]
perf_pref[M] AND dur_call[H]
    --> [PFI[L]@0.50%, SHAP[L]@0.50%, Anchors[VL]@0.50%, CFProto[VL]@0.50%]
perf_pref[M] AND num_feat[M]
    --> [Anchors[L]@0.50%]
perf_pref[M] AND num_feat[H]
    --> [PFI[L]@0.50%, SHAP[VL]@0.50%, Anchors[VL]@0.50%, CFProto[L]@0.50%]
perf_pref[H] AND dur_call[M]
    --> [Anchors[L], CFProto[VL]]
perf_pref[H] AND dur_call[H]
    --> [PFI[L], SHAP[L], Anchors[VL], CFProto[VL]]
perf_pref[H] AND num_feat[M]
    --> [Anchors[L]]
perf_pref[H] AND num_feat[H]
    --> [PFI[L], SHAP[VL], Anchors[VL], CFProto[L]]
ordinal_feat[True] AND corr[VL]
    --> [CFProto[VL]@0.60%]
ordinal_feat[True] AND corr[L]
    --> [CFProto[L]@0.60%]
ordinal_feat[True] AND discr[L]
    --> [CFProto[VL]@0.80%]
ordinal_feat[True] AND discr[M]
    --> [CFProto[L]@0.60%]
prep_time[L] --> [PFI[L], SHAP[L], Anchors[VL], CFProto[VL]]
prep_time[M] --> [Anchors[L], CFProto[L]]
scope_global[True]
    --> [PDP + ICE[VH], ALE[VH], PFI[VH], SHAP[VH]]
scope_local[True]
    --> [PDP + ICE[VH], SHAP[VH], Anchors[VH], CFProto[VH]]

```

## A.8. Screenshots der GUI der Web-Anwendung

Anmerkung:

Die nachfolgenden Screenshots dienen nur der Visualisierung des Aufbaus der einzelnen Webanwendungsseiten, weshalb die Lesbarkeit der Texte hier vernachlässigt wurde.

Der XAIR ist unter <https://github.com/viadee/xair> für eine eigene Verwendung bzw. Bereitstellung verfügbar.

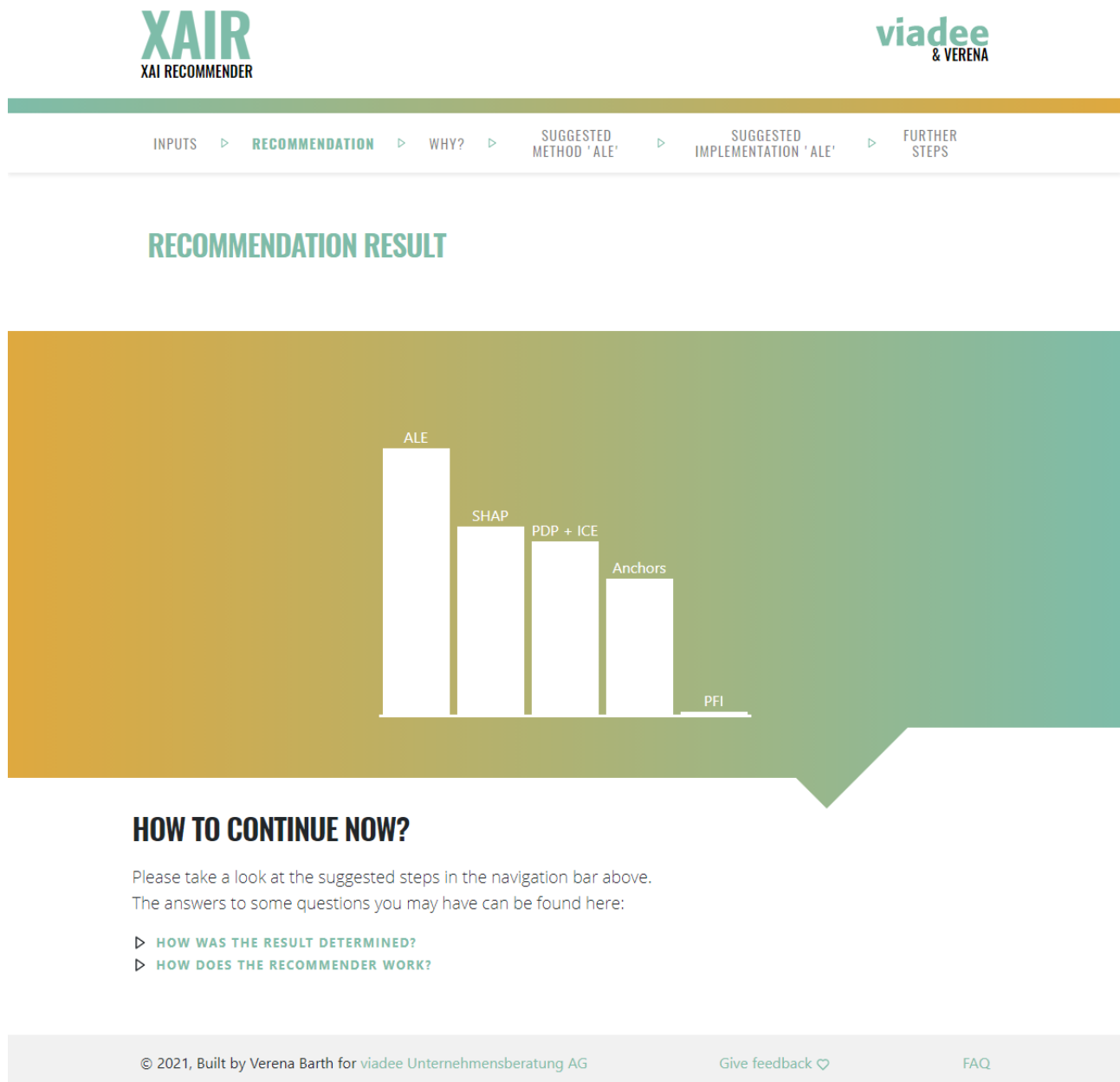


Abbildung A.7.: Screenshot Ergebnisseite



## RECOMMENDATION OF A SUITABLE XAI METHOD

### INSERT VALUES FOR GETTING A RECOMMENDATION

In order to be able to propose the most appropriate XAI method for your ML model, you first of all have to provide some properties of your ML model and its underlying data. Additionally you can add personal preferences, for example regarding the method's explanation type or it's usage.

- ▷ WHY USING XAI?
- ▷ HOW DOES IT WORK?

### MODEL SPECIFIC INPUT PARAMETERS

Model name	<input type="text" value="Model Titanic"/>	<a href="#">?</a>
<input checked="" type="checkbox"/> Model available		<a href="#">?</a>
<input checked="" type="checkbox"/> Classification task		<a href="#">?</a>
<input type="checkbox"/> Class probabilities		<a href="#">?</a>
<input checked="" type="checkbox"/> Preprocessing Operations available		<a href="#">?</a>
Duration of Model Call	short <input type="range" value="25"/> long	<a href="#">?</a>

### DATA SPECIFIC INPUT PARAMETERS

The suitability of an XAI method highly depends on the nature of the training data. Therefore, some properties of these must be obtained and considered for a recommendation. Several options are available for the specification: An approximate specification via the tab "Approximate Input" is usually already sufficient. However, if the required parameters are already available, you can specify them in the "Exact Input" tab. Unfortunately, the option of automatic data analysis is not yet available.

- ▷ WHICH FORMAT OF INPUT DATA SHOULD BE TAKEN INTO CONSIDERATION?

Approximate input	Exact input	Automatic input by file upload
-------------------	-------------	--------------------------------

### APPROXIMATE INPUT

Correlation	negligible	<input type="range" value="50"/>	high	<a href="#">?</a>
Discretizability	poor	<input type="range" value="25"/>	good	<a href="#">?</a>
Number of Features	few	<input type="range" value="25"/>	many	<a href="#">?</a>
Features of Interest (FOI)	<input type="text" value="Sex, Age"/>			<a href="#">?</a>
Correlation FOI	negligible	<input type="range" value="75"/>	high	<a href="#">?</a>
Discretizability FOI	poor	<input type="range" value="50"/>	good	<a href="#">?</a>
<input checked="" type="checkbox"/> Labels available				<a href="#">?</a>
<input checked="" type="checkbox"/> Ordinal Features				<a href="#">?</a>

### USER PREFERENCES REGARDING XAI METHOD

Performance Preference	doesn't matter	<input type="range" value="50"/>	high	<a href="#">?</a>
Preparation Time Preference	fast	<input type="range" value="50"/>	doesn't matter	<a href="#">?</a>
<input checked="" type="checkbox"/> Local Scope (optional)				<a href="#">?</a>
<input type="checkbox"/> Global Scope (optional)				<a href="#">?</a>

SUBMIT

Abbildung A.8.: Screenshot Eingabeseite

INPUTS ▶ RECOMMENDATION ▶ **WHY?** ▶ SUGGESTED METHOD 'ALE' ▶ SUGGESTED IMPLEMENTATION 'ALE' ▶ FURTHER STEPS

## HOW DID THE RESULTS COME?

### WHAT INPUTS WERE USED TO DETERMINE THE SUITABILITY OF THE XAI METHODS?

**BOOLEAN CRITERIA**

- ☒ Classification task
- ☒ Labels available
- ☒ Model available
- ☒ Ordinal Features
- ☒ Class probabilities
- ☒ Preprocessing Operations available
- ☒ Global Scope (optional)
- ☒ Local Scope (optional)

**OTHER CRITERIA**

- Correlation medium strong
- Correlation FOI strong/very strong
- Discretizability bad/medium
- Discretizability FOI medium/good
- Duration of Model Call fast/medium
- Number of Features few/medium
- Ordinal Features yes
- Performance Preference low/medium
- Preparation Time Preference medium
- Global Scope (optional) no
- Local Scope (optional) yes

### WHICH METHODS CAN NOT BE USED?

CFProto: ☒ Class probabilities

### HOW WAS THE SUITABILITY OF THE XAI METHODS EVALUATED FOR 'MODEL TITANIC'?

Because of ...	PDP + ICE is	ALE is	PFI is	SHAP is	Anchors is
Correlation FOI: <b>strong</b>	very bad	very good	bad	bad	rather bad
Local Scope: <b>yes</b>	very good			very good	very good
Preparation Time Preference: medium					bad
Discretizability: <b>medium</b>					bad
Correlation: <b>medium strong</b>			bad	bad	rather bad
Performance Preference: medium + Number of Features: medium					rather bad
Performance Preference: medium + Duration of Model Call: <b>medium</b>					rather bad
Discretizability FOI: <b>medium</b>					bad
Correlation FOI: <b>strong</b> + Discretizability FOI: <b>medium</b>		good			
Discretizability: <b>bad</b>		rather bad			very bad
Correlation FOI: <b>strong</b> + Discretizability FOI: <b>good</b>		very good			

Note: Some of the criteria listed here may have different values, for example "medium" and "high". This is because the given input value cannot be completely assigned to any term. The characteristics of the inputs are listed **above**.

### WHICH XAI METHODS WERE TAKEN INTO CONSIDERATION?

ACCUMULATED LOCAL EFFECTS

ANCHORS

COUNTERFACTUALS GUIDED BY PROTOTYPES

PARTIAL DEPENDENCE PLOTS (PDP) & INDIVIDUAL CONDITIONAL EXPECTATION (ICE)

PERMUTATION FEATURE IMPORTANCE

SHAPLEY ADDITIVE EXPLANATIONS

© 2021, Built by Verena Barth for viadee Unternehmensberatung AG

Give feedback

FAQ

Abbildung A.9.: Screenshot Erklärungsseite

**XAIR**  
XAI RECOMMENDER

**viadee**  
& VERENA

INPUTS > RECOMMENDATION > WHY? > **SUGGESTED METHOD 'ALE'** > SUGGESTED IMPLEMENTATION 'ALE' > FURTHER STEPS

## ALE (ACCUMULATED LOCAL EFFECTS)

### TAXONOMISCHE EINORDNUNG

GLOBALE ERKLÄRUNG

FEATURE RELEVANZ

PERTURBATIONSBASTIERT

VISUELLE ERKLÄRUNG

> WELCHE IMPLEMENTIERUNG WIRD EMPFOHLEN?

### WELCHE FRAGE BEANTWORTET ALE?

Was sagt das Modell vorher, wenn man das Feature innerhalb eines kleinen Intervalls um seinen Wert ändert?

*Welchen durchschnittlichen Einfluss hat das Alter auf die Bewilligung eines Kredits durch die Bank?*

### WIE FUNKTIONIERT ALE?

Für die Berechnung des Einflusses des Features wird dessen Wertebereich zuerst in lokale Bereiche unterteilt. Für jeden Bereich wird der Feature-Wert zwischen den Rändern bewegt und der durchschnittliche Unterschied der Vorhersage berechnet: „Was sagt das Modell vorher, wenn man das Feature innerhalb eines kleinen Intervalls um seinen Wert ändert?“ [1]. Eine Lokalisierung der Feature-Werte mithilfe dieser Bereiche vermeidet eine Extrapolation und somit die Erstellung unrealistischer Dateninstanzen aufgrund von Korrelation der Eingabe-Features [2]. Die Summe dieser arithmetischen Mittel wird zentriert dargestellt (Mittelwert = 0). Die Berechnung des Einflusses zweier Features ist auch möglich. Der Graph zeigt aber im Gegensatz zu **PDP Plots** ausschließlich den Effekt zweiter Ordnung, d.h. die zusätzliche Auswirkung der Interaktion der beiden Features auf die Vorhersage, an. [1]

### WAS IST DAS ERGEBNIS VON ALE?

Die Visualisierung zeigt somit die (globale) Auswirkung des Feature-Werts auf die Vorhersage im Vergleich zu der durchschnittlichen Vorhersage. [1]

Ergebnisdarstellung ALE

### WELCHE FRAGE BEANTWORTET ALE NICHT?

2D ALE Plots (Plots für zwei Features) zeigen ausschließlich den Effekt zweiter Ordnung, d.h. die zusätzliche Auswirkung der Interaktion der beiden Features auf die Vorhersage an, nicht den Gesamteffekt. Wenn zwei Features nicht interagieren, aber beide einen linearen Effekt auf die Vorhersage haben, ist die 1D ALE Kurve beider eine Gerade an, die 2D ALE Kurve ist nahe 0, da es durch Featureunabhängigkeit keinen zusätzlichen Interaktionseffekt gibt. [1]

### WELCHE IMPLEMENTIERUNG WIRD EMPFOHLEN?

### WEITERE METHODEN

ANCHORS

COUNTERFACTUALS GUIDED  
BY PROTOTYPES

PARTIAL DEPENDENCE PLOTS  
(PDP) & INDIVIDUAL  
CONDITIONAL EXPECTATION  
(ICE)

PERMUTATION FEATURE IMPORTANCE

SHAPLEY ADDITIVE EXPLANATIONS

### REFERENZEN

[1] Molnar, Christoph (2020): Interpretable Machine Learning

[2] Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. (2014), 'Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation'.

© 2021, Built by Verena Barth for viadee Unternehmensberatung AG

Give feedback

FAQ

Abbildung A.10.: Screenshot Detailseite der empfohlenen XAI-Methode (ALE)

[INPUTS](#)
[RECOMMENDATION](#)
[WHY?](#)
[SUGGESTED METHOD 'ALE'](#)
[SUGGESTED IMPLEMENTATION 'ALE'](#)
[FURTHER STEPS](#)

## IMPLEMENTATION FOR 'ALE'

### WELCHE IMPLEMENTIERUNG WIRD EMPFOHLEN?

Die Auswahl guter Python-Implementierungen für ALE ist klein, die Unterstützung durch eine große Community selten gegeben und die Dokumentation oft unzureichend. Trotz einer kleinen Community auf GitHub fällt auf die Entscheidung auf **PyALE**, da sie als einzige die Visualisierung kategorischer Features unterstützt.

## ALE IMPLEMENTIERUNG: PYALE

[DOKUMENTATION](#)
[SOURCECODE](#)

### WAS BENÖTIGT MAN DAFÜR?

- Zugriff auf Modell
- Trainingsdaten
  - Kategorische Features: Kodiert wenn Label Encoded, nicht kodiert (original Feature-Werte) bei angewendetem One-Hot-Encoding (OHE)
  - Numerische Features: Nicht skaliert
  - Zugriff auf Spaltennamen: Benötigt
- Bei One-Hot-Encoding (OHE) wird die Kodierungsfunktion benötigt.

### WAS IST DAS ERGEBNIS?

- 1D ALE Plots für jeweils ein numerisches oder kategorielles Feature, inklusive optionaler Darstellung eines Konfidenzintervalls
- 2D ALE Plots für numerische Features

!

### WAS MUSS MAN BEACHTEN?

🔧

### HINWEISE ZUR ANWENDUNG DER METHODE

- Für die Definition der Bereiche werden die Quantile der Feature Verteilung verwendet. Durch die Verwendung der Quantile wird sichergestellt, dass sich in jedem der Bereiche gleich viele Dateninstanzen befinden. Quantile können allerdings sehr unterschiedlich große Bereiche haben, was bei einem sehr ungleich verteilten Feature zu seltsamen ALE-Diagrammen führen kann. Daher sollten Outliers, einzelne Dateninstanzen mit unüblichen, extrem hohen oder niedrigen Featurewerten, aus den Trainingsdaten entfernt werden.
- ALE Plots brauchen eine natürliche Ordnung der Featurewerte. Sollten kategorielle Features keine Ordnung aufweisen, werden die Kategorien nach ihrer Ähnlichkeit auf Basis der anderen Features geordnet. Die Distanz zweier Kategorien ist die Summe der Distanzen aller anderen Features (**empirical cumulative distribution function**). [1]
- 2D ALE Plots können das Fehlverhalten des Modells durch Eingabe sehr unwahrscheinlicher bzw. unmöglicher Dateninstanzen aufgrund der Extrapolation (dem Verlassen des 'normalen' Wertebereiches eines Features) sichtbar machen. Es ist eine Designentscheidung, ob dieser Bereich unwahrscheinlicher Dateninstanzen einbezogen werden sollte oder nicht. Das kann abhängig von der Tatsache sein, ob die Verteilung der Testdaten oder neuer, während dem Einsatz des Modells erhaltener Daten, anders erwartet wird. [2]
- 2D ALE Plots eignen sich, wenn Interesse an der Feature Interaktion besteht. Wenn Interesse am kombinierten Effekt der Merkmale, d.h. am Gesamteffekt und den Interaktionseffekten besteht, dann sollten PDPs bevorzugt werden.
- Für diesen XAI Methoden Vorschlag wurde hauptsächlich die Korrelationen der angegebenen Features of Interest betrachtet. Sollten andere, nicht korrelierende Features ebenfalls visualisiert werden und die Berechnungszeit irrelevant sein, sind PDP Plots mit zusätzlichen ICE Kurven durch ihre einfache Verständlichkeit zu bevorzugen.
- Viele Beispiele für die Anwendung von PyALE sind [hier](#) zu finden

</>

### HINWEISE ZUR PARAMETERAUSWAHL FÜR DIE IMPLEMENTIERUNG

- `grid_size`  
Definiert die Anzahl der lokalen Bereiche: Eine zu hohe Anzahl kann zu einer wackeligen Kurve mit vielen kleinen Auf- und Abschwüngen führen. Wenige glätten das Diagramm, machen es somit aber auch ungenau und eventuell bleibt die tatsächliche Komplexität des Modells verborgen.
- `include_CI` und `C`  
Die Methode erlaubt eine Darstellung eines Konfidenzintervalls mithilfe von Zufallsstichproben des Datensatzes um den geschätzten Feature Effekt.

📖

### REFERENZEN

[1] Jomar, Dana (2020): PyALE  
 [2] Molnar, Christoph (2020): Interpretable Machine Learning

© 2021, Built by Verena Barth for viadee Unternehmensberatung AG

[Give feedback](#)

[FAQ](#)

Abbildung A.11.: Screenshot Detailseite der empfohlenen Implementierung (ALE)

INPUTS
RECOMMENDATION
WHY?
SUGGESTED METHOD 'ALE'
SUGGESTED IMPLEMENTATION 'ALE'
FURTHER STEPS

## RECOMMENDATIONS FOR FURTHER STEPS

Um ein erklärbares, nachvollziehbares und faires Modell zu erhalten bzw. zu gewährleisten, empfehlen wir folgende Punkte zu beachten:

- Es sollte eine Visualisierungsmethode für beliebig viele, jedoch für mindestens die kritischen Features (FOI) verwendet werden, sodass ihre Auswirkungen auf das Ergebnis genauer betrachtet werden können.
- Um ein besseres Gesamtbild des Modells und eine vollständigere Erklärung zu erhalten, ist neben der Anwendung der empfohlenen Methode die mindestens einer weiteren XAI Methode ratsam. Eine Kombination von lokalen und globalen Methoden ist empfehlenswert, wobei nach dem Local-first Ansatz gehandelt werden sollte [1]. Das Ausführen mehrerer Methoden ist wichtig, da bestimmte XAI Methoden angegriffen und getäuscht werden können, siehe [2]. Unten ist eine Liste weiterer empfohlener XAI Methoden zu finden.
- Lokale Erklärungen sollten dabei sowohl für richtig als auch für falsch vorhergesagte Dateninstanzen erzeugt werden; das gibt Aufschlüsse über eine mögliche Verbesserung des Feature Engineerings. [3]
- Es ist wichtig, dass die Modellperformance nur in Kenntnisnahme einer globalen Feature Importance Methode betrachtet wird. [3]
- Auch wenn das Modell in Betrieb ist sollten die Eingabedaten kontinuierlich auf Änderungen geprüft und überwacht werden, um neuere Entwicklungen widerzuspiegeln, das Modell zu aktualisieren und somit Vertrauen in die Robustheit des Modells zu gewährleisten. [3]

## OTHER RECOMMENDED XAI METHODS

### 1 SHAP (SHAPLEY ADDITIVE EXPLANATIONS)

LOKALE UND GLOBALE ERKLÄRUNG
FEATURE RELEVANZ
PERTURBATIONSBASIERT

Wie hoch ist der Beitrag jedes Merkmals (oder einer Gruppe von Merkmalen) zur Vorhersage einer Dateninstanz; verglichen mit der durchschnittlichen Vorhersage für alle Instanzen?

### 2 PDP + ICE (PARTIAL DEPENDENCE PLOTS (PDP) & INDIVIDUAL CONDITIONAL EXPECTATION (ICE))

LOKALE UND GLOBALE ERKLÄRUNG
FEATURE RELEVANZ
PERTURBATIONSBASIERT
VISUELLE ERKLÄRUNG

PDP: Wie ist der durchschnittliche Zusammenhang zwischen dem betrachteten (von den anderen unabhängigen) Feature und der Vorhersage?  
ICE: Wie ist der dateninstanzspezifische Zusammenhang zwischen dem betrachteten (von den anderen unabhängigen) Feature und der Vorhersage?

### 3 ANCHORS

LOKALE ERKLÄRUNG
MODELLVEREINFACHUNG
PERTURBATIONSBASIERT

Nach welchen Regeln lässt sich die Vorhersage einer Instanz erklären? Auf wie viele Instanzen trifft die Regel zu (Coverage) und wie gut beschreibt sie die Vorhersage der Instanz (Precision)?

### 4 PFI (PERMUTATION FEATURE IMPORTANCE)

GLOBALE ERKLÄRUNG
FEATURE RELEVANZ
PERMUTATIONSBASIERT



Wie viel Einfluss hat das Feature auf die Korrektheit der Vorhersage, d.h. wie verändert sich der Vorhersagefehler des Modells nach Veränderung des Feature-Wertes?

## REFERENZEN

- [1] Leslie, David (2019): Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute.
- [2] Dylan Slack; Sophie Hilgard; Emily Jia; Sameer Singh; Himabindu Lakkaraju (2020): Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods.
- [3] Kangur, Ayla (2020): Explainable AI In practice. How committing to transparency made us deliver better AI products. In: Towards Data Science, 24.12.2020.
- [4] Molnar, Christoph (2020): Interpretable Machine Learning.



© 2021, Built by Verena Barth for viadee Unternehmensberatung AG
Give feedback
FAQ



Abbildung A.12.: Screenshot Allgemeine Empfehlungsseite

INPUTS > RECOMMENDATION > WHY? > SUGGESTED METHOD 'ALE' > SUGGESTED IMPLEMENTATION 'ALE' > FURTHER STEPS

## FAQ


**WHY SHOULD I USE EXPLAINABLE AI (XAI)?**



**WHY SHOULD I USE THIS XAI RECOMMENDER?**


Recent XAI research has presented a variety of XAI methods and implementations in the form of stand-alone prototype solutions. However, these are hardly used in practice. This is partly because XAI is a new and rapidly developing field, and partly because the existing knowledge is scattered and needs to be organized.

Some scientific publications classify diverse methods hierarchically, but rarely provide concrete guidelines or advices for their application. Government agencies such as the Federal Office for Information Security (BSI) point out the relevance of explainability of AI systems and that method-specific properties of inputs must be considered in the selection process, and plausible explanations must be enabled ([Federal Office for Information Security 2021, p. 41](#)). However, it does not provide information to help users successfully select and apply an XAI method that is appropriate in this sense.



So, not all XAI methods are ideally suited in all contexts of use. For example, results of perturbation-based methods can be falsely influenced by correlations in the input features.



The **XAI Recommender** considers all suitability-influencing parameters and suggests a suitable method for your data and model context. For the application, only vague knowledge about the training dataset and the context of use is assumed.



In addition to a justified method suggestion, it also provides an explanation of the method, a recommended implementation and hints for the application. Thus, you not only get support for the (quick) selection, but also for the application! Furthermore, information about all methods in this system will be provided.



**SHORT SUMMARY OF BENEFITS USING THE XAI RECOMMENDER**



- Get a justified recommendation of an applicable XAI method
- Learn which XAI method should be used under which circumstances
- Get a suggested implementation with some hints for its application
- Inform yourself about all available XAI methods
- Learn German (sorry, information pages are currently not available in English)


**WHAT DOES IT MEAN IF A XAI METHOD IS SUITABLE?**



**HOW DOES THE XAI RECOMMENDATION SYSTEM WORK?**



**WHICH XAI METHODS WERE TAKEN INTO CONSIDERATION?**



**WHICH FORMAT OF INPUT DATA SHOULD BE TAKEN INTO CONSIDERATION?**



**WHAT DOES THE TERM "DISCRETIZABILITY" MEAN?**


© 2021, Built by Verena Barth for viadee Unternehmensberatung AG
Give feedback
FAQ

Abbildung A.13.: Screenshot FAQ-Seite

## A.9. Anwendungsfall der Evaluation

### Klassifikationsaufgabe anhand des Titanic Datensatzes: Wer überlebt?

Du hast Zugriff auf das zu erklärende Modell und die einzelnen Datentransformationsschritte des Preprocessings sind ebenfalls verfügbar. Die Klassenwahrscheinlichkeiten, d.h. wie wahrscheinlich das Überleben bzw. Sterben eines Titanic-Gastes ist, werden nicht zurückgegeben. Der Modellaufruf und der Erhalt der Vorhersage ist sehr schnell.

Von Deinem Team hast Du einen ausführlichen Report zur Einschätzung des Datensatzes erhalten. Aus diesem sind die datenspezifischen Eingaben für das XAIR nach deiner Einschätzung vorzunehmen. Tipp bezüglich der Korrelationen: Der Phik Korrelationskoeffizient kann die paarweisen Korrelationskoeffizienten für kategoriale, ordinale und Intervall-Features berechnen und ist daher den anderen, variabeltypabhängigen Korrelationskoeffizienten zu bevorzugen.

Nun liegt es an Dir eine passende XAI-Methode für das Modell auszuwählen, die sich für den gegebenen Daten- und Nutzungskontext und Deine Präferenzen eignet! Verwende dafür bitten den XAIR, der unter der folgenden URL verfügbar ist:

`http://xairecommender-frontend.germanywestcentral.azurecontainer.io/`

Bitte investiere mindestens 10 Minuten in die Verwendung des Systems. Du kannst Dir die empfohlene Methode genauer ansehen, Dich über andere Methoden informieren, ggf. abwägen, ausprobieren, rumklicken etc. ...

## A.10. Leitfaden des Fragenkatalogs zur Evaluation

Anmerkungen:

- *Das Interview findet mit denselben Personen statt, die bereits an der Anforderungsanalyse teilgenommen haben*
- *Vor dem Interview holen sich die Interviewteilnehmer eigenständig und unabhängig eine Empfehlung für den in Anhang A.9 aufgeführten, konstruierten Anwendungsfall ein*
- *Keiner der Personen setzte sich vor der Beantwortung näher mit dem Inhalt der Methoden- und Implementierungsdetailseiten auseinander*

### Kategorie 1: Erwartungen

1. Warum oder mit welchem Ziel würdest Du das XAI-Empfehlungssystems nutzen?
2. Welche Erwartungen hattest Du an das Empfehlungssystem?
3. Inwieweit wurden Deine Erwartungen erfüllt bzw. NICHT erfüllt?
4. Wie viel Zeit würdest Du für einen Empfehlungserhalt investieren, wenn eine XAI-Methode für das Modell angewendet werden soll/muss?

### Kategorie 2: Nutzen

1. War das Deine erste Anwendung des in dieser Weise aufgebauten XAI-Empfehlungssystems?
  - a) Ist externe Hilfe für die Anwendung erforderlich?
  - b) Kann man sich die Bedienung nach einer längeren Nicht-Nutzung merken?
2. Wie viel Erfahrung in ML/XAI wird Deiner Meinung nach für die Verwendung des XAI-Empfehlungssystems vorausgesetzt?
3. Wie detailliert waren Deine Kenntnisse über die Gegebenheiten bei Angabe der Eingabeparameter?
  - a) Ist das Voraussetzen dieser Kenntnisse legitim?
4. Hast Du die von Dir gewollten/geforderten Informationen schnell gefunden?
  - a) Gibt es Informationen, die Du vermisst?
5. Konnte es Dir Einblicke in XAI und vorhandene XAI-Methoden bieten?
6. Hast Du Wissen erlangen können, auf das Du für kommende XAI Auswahlentscheidungen zurückgreifen kannst?

### <Vorstellung des XAI-Empfehlungssystems>

### Kategorie 3: Verwendbarkeit und Qualität

1. Wie erleichtert Dir das XAIR die Auswahl einer XAI-Methode?
2. Kennst Du die vorgeschlagene Methode?



- a) Ist sie für Deinen Use Case geeignet?
- b) Stimmt sie mit Deiner ursprünglich vorgesehenen Methode überein?
- 3. Bietet es Dir hinsichtlich der XAI-Methode genug Information um sie
  - a) zu verstehen
  - b) anzuwenden
- 4. Kannst Du nachvollziehen, wie und warum die Empfehlung generiert wurde?
- 5. „Vertraust“ Du dem XAI-Empfehlungssystem?
  - a) Was kann man tun, um Vertrauen zu erhöhen?

#### **Kategorie 4: Systemoberfläche**

- 1. Wie intuitiv ist die Struktur/der Aufbau des XAI-Empfehlungssystems?
- 2. Sind die Navigation und möglichen Funktionen intuitiv verständlich?
- 3. Sind die Inhalte/Texte der Hilfestellung etc. leicht verständlich?
  - a) Wie vertraut bist Du mit Termini des ML/XAI?
  - b) Ist die verwendete Sprache adäquat?
- 4. Wie findest Du den Aufbau und das Design der Web-Anwendung?

#### **Kategorie 5: Motivation/Produktivität**

- 1. Ist der (zeitliche) Aufwand für die Nutzung des Systems akzeptabel?
  - a) Aufwand der Parameterangabe
  - b) Dauer des Ergebniserhalts
  - c) Zeitinvestition zum Verstehen des Empfehlungsergebnisses
- 2. Ist der (zeitliche) Aufwand für den Erhalt des Empfehlungsergebnisses gerechtfertigt?
- 3. Bist Du motiviert, die XAI-Methode anzuwenden/anwenden zu lassen?
- 4. Würdest Du das XAI-Empfehlungssystem nochmals/mehrmals nutzen?
- 5. Würdest Du es Deinen Kollegen empfehlen?

#### **Kategorie 5: Weiteres Feedback**

Hast Du weitere Verbesserungsvorschläge, Anregungen, Wünsche?