viadee®
IT-Unternehmensberatung

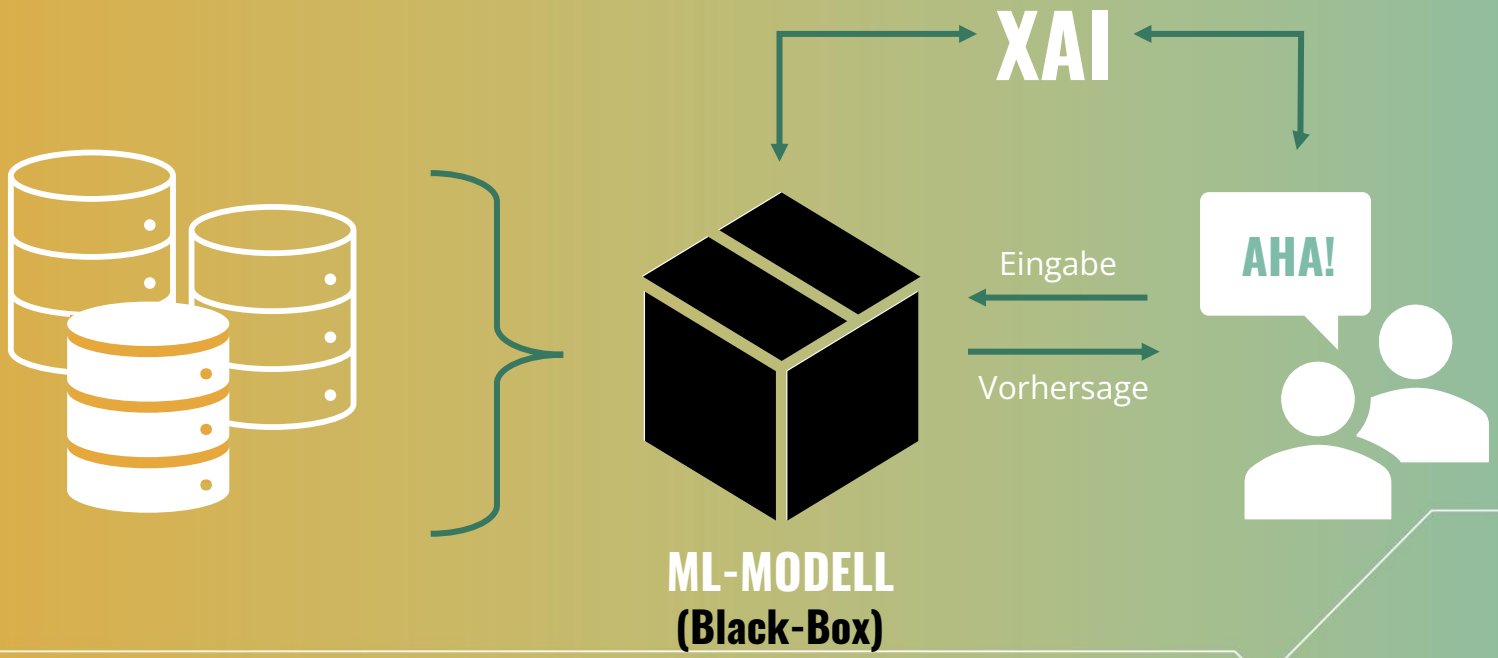
EXPLAINABLE AI VERFAHREN

und die Herausforderung ihrer Anwendung

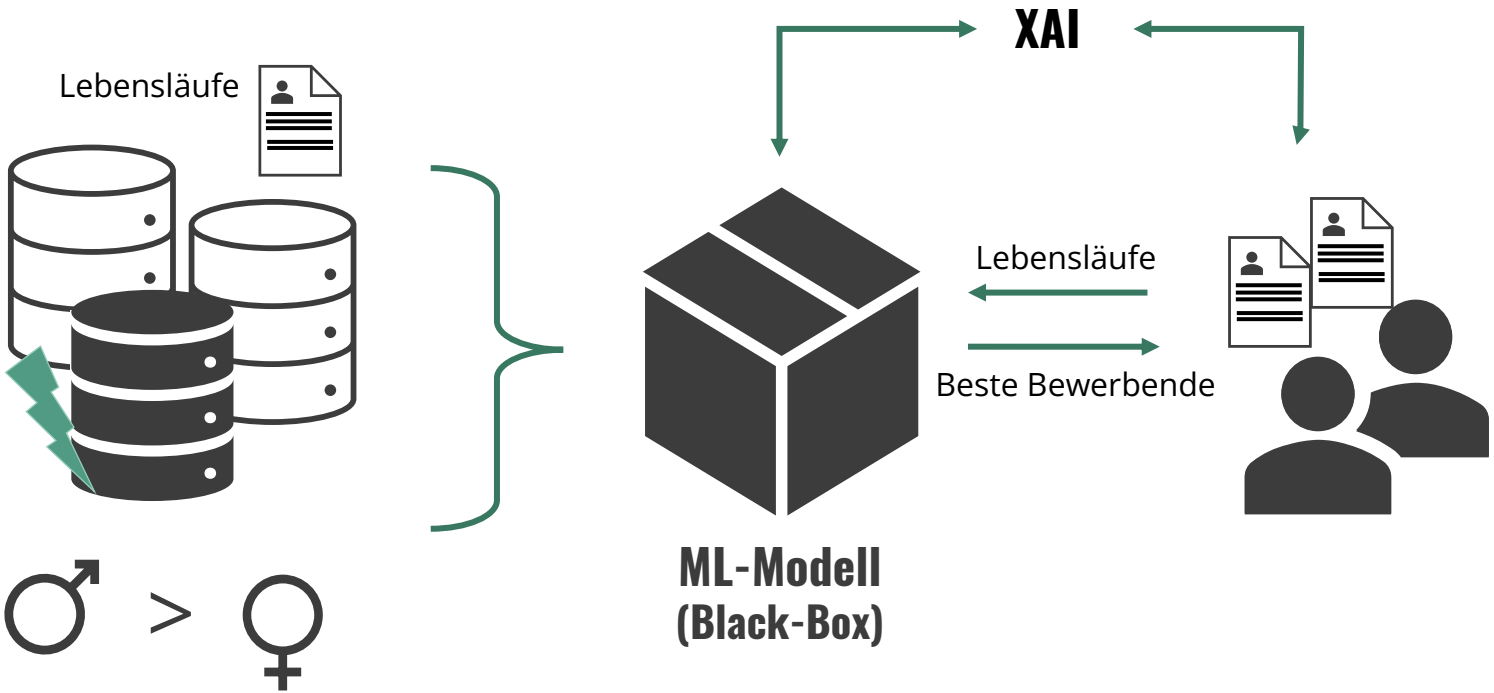
Verena Barth

viadee®
IT-Unternehmensberatung

WARUM XAI?



AMAZON: AUTOMATISIERTER EINSTELLUNGSPROZESS [1]



PROBLEMANALYSE

WARUM WIRD XAI NICHT ANGEWENDET?

- XAI ist ein neuer, sich rasant entwickelnder Bereich [1]
- Wissen ist verstreut und unorganisiert [2]

→ Keine Unterstützung bei der Auswahl und Anwendung geeigneter Methoden
- Sehr anwendungsfallspezifisch
- Viele offene Fragen (z.B. Metrik für Erklärbarkeit) [3]

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

WARUM WIRD XAI NICHT ANGEWENDET?

- XAI ist ein neuer, sich rasant entwickelnder Bereich [1]
- Wissen ist verstreut und unorganisiert [2]

→ Keine Unterstützung bei der Auswahl und Anwendung geeigneter Methoden
- Sehr anwendungsfallspezifisch
- Viele offene Fragen (z.B. Metrik für Erklärbarkeit) [3]

→ **XAI EMPFEHLUNGSSYSTEM (XAIR)**

ZIEL

XAI-EMPFEHLUNGSSYSTEM (XAIR)

viadee®
IT-Unternehmensberatung

- Auswahl der für den Anwendungskontext geeigneten XAI-Methoden erleichtern
- Förderung der tatsächlichen Anwendung empfohlener XAI-Methoden
- Empfehlung jederzeit und ohne viel Aufwand einholbar
- Nachschlagewerk

ANNAHMEN

- Empfehlung für existierendes ML-Modell
- Klassifikation/Regression mit tabellarischen Daten
- Zielgruppe: Modellentwickler und –verantwortliche (ohne tieferen ML-Kenntnisse)

viadee®
IT-Unternehmensberatung

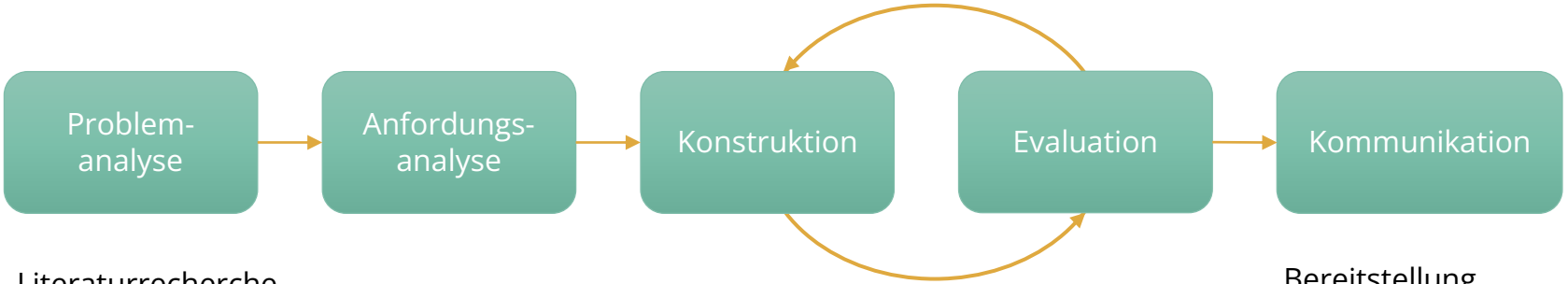
AGENDA

- Methodik
- Konzeption des XAIR
  - Eignungsbeurteilung einer XAI-Methode
  - Eignungsbeeinflussende Kriterien
  - Formalisierung des Wissens
  - Konfiguration des Expertensystems
- Live-Demo des XAIR
- Evaluation
  - Nutzerbefragung
  - Limitationen und Herausforderungen
  - Einordnung in den Forschungsstand

viadee®
IT-Unternehmensberatung

VERWENDETE METHODIK

Requirements Driven Design Science Research Ansatz [4]



Literaturrecherche nach [5]

Bereitstellung
- Empfehlungssystem
- Sourcecode

Anforderungsanalyse und Evaluation:
Halbstrukturierte Online-Einzelinterviews mit 5 Personen der Zielgruppe

viadee®
IT-Unternehmensberatung

KONZEPTION EINES XAI-EMPFEHLUNGSSYSTEMS

EIGNUNG EINER XAI-METHODE

Ergebnisbeurteilung nicht sinnvoll

→ Erklärung subjektiv, kontextabhängig [6]

Beurteilung anhand von Eigenschaften, die

- Methodenanwendung erschweren/unmöglich machen
- Negativen/verfälschenden Einfluss auf Ergebnis haben
- Interpretierbarkeit der Erklärung mindern/verkomplizieren

KONZEPTION

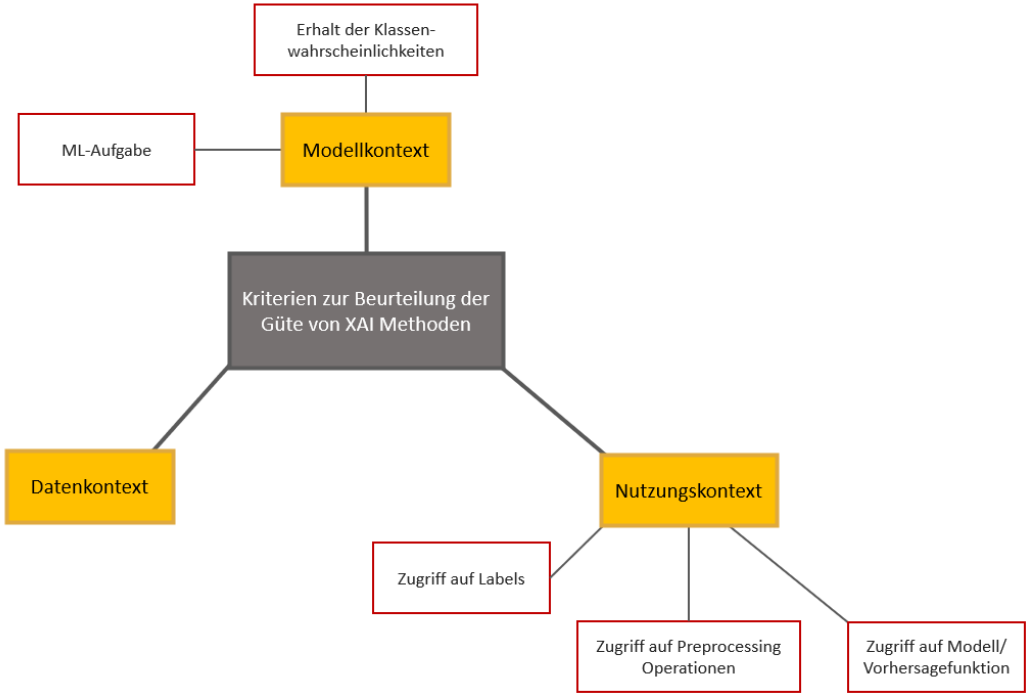
METHODENAUSWAHL

viadee®
IT-Unternehmensberatung

VISUALISIERUNG	**PDP + ICE** Partial Dependence Plot + Individual Conditional Expectation	**ALE** Accumulated Local Effects
FEATURE RELEVANZ	**SHAP** SHapley Additive exPlanations	**PFI** Permutation Feature Importance
BEISPIELBASIERT	**CFProto** Counterfactuals guided by Prototypes	
MODELLVEREINFACHUNG	**Anchors**	

Lokale Erklärung Globale Erklärung Globale und lokale Erklärung

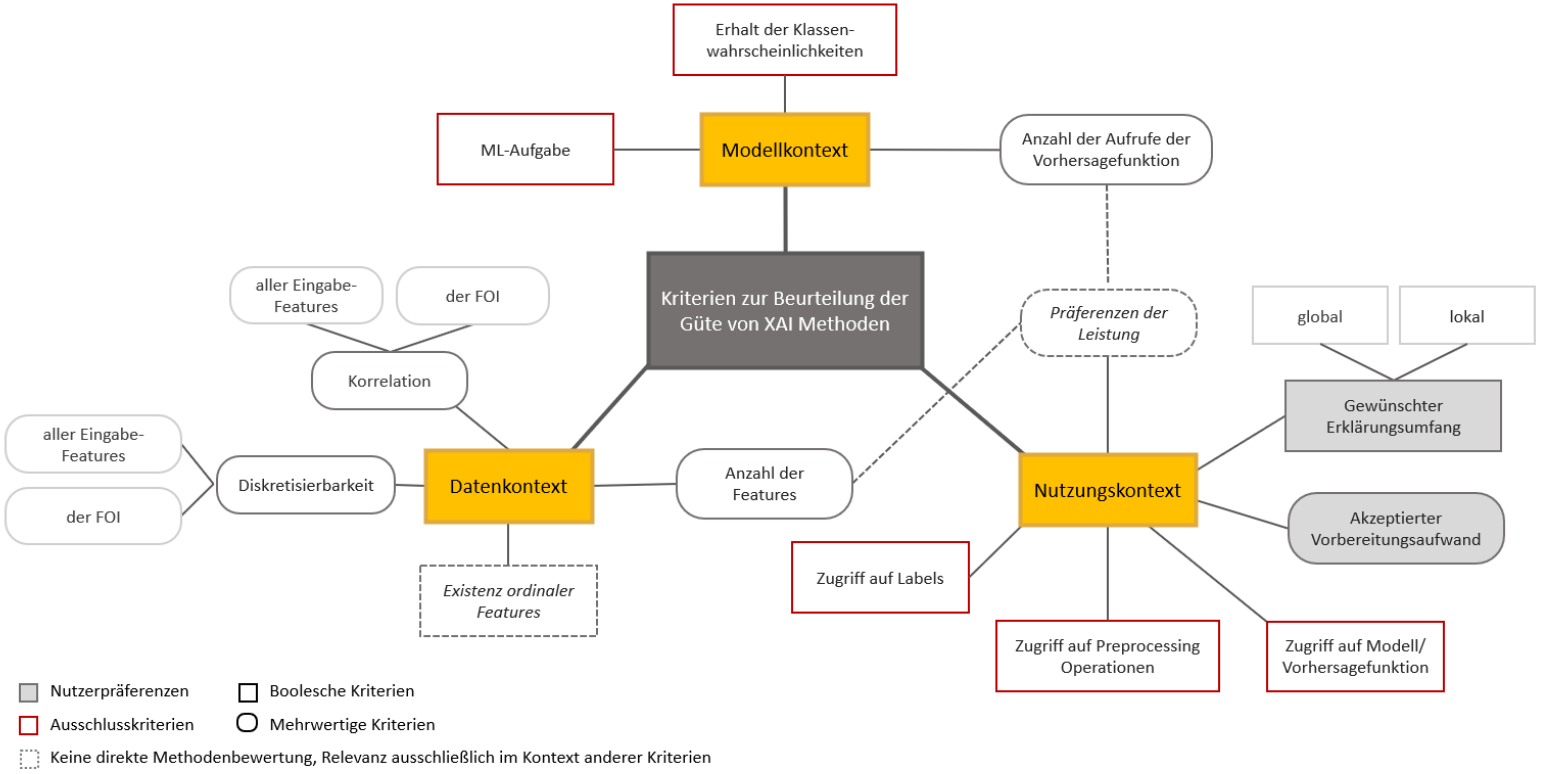
METHODENVORAUSSETZUNGEN



Boolesche Kriterien

Ausschlusskriterien

EIGNUNGSBEEINFLUSSENDE KRITERIEN



PROBLEM DER UNGENAUIGKEIT

- Bei Erhebung der Kriterienwerte und der Auswahl einer Wissensrepräsentation

- Keine Schwellwerte
  - Bspw.: Wie kann man die Korrelation eines Datensatzes messen?
    → keine exakte Quantifizierung möglich
- Teilweise subjektiv oder nicht-deterministisch
  - Bspw.: Wann hat ein Datensatz viele Features?
- Vage Aussagen bzgl. Beeinflussung der Methodeneignung

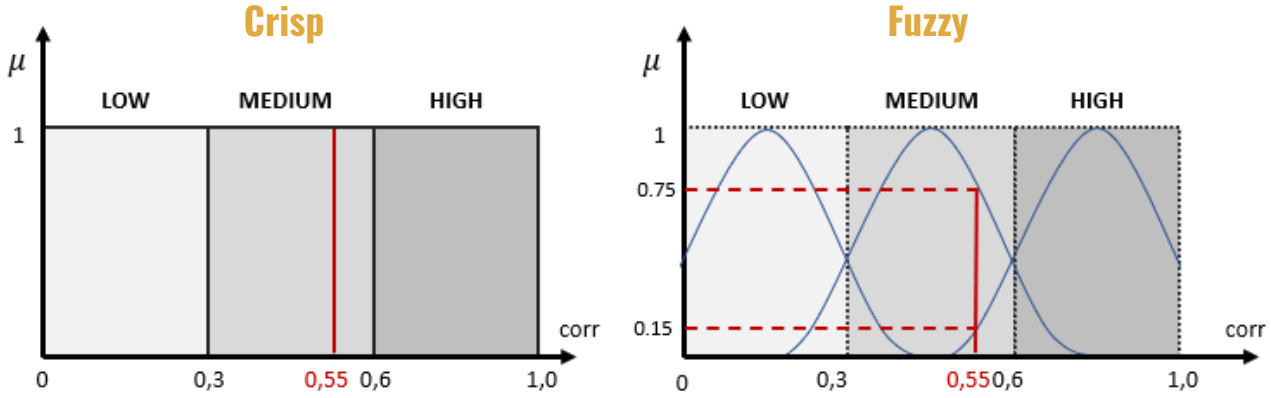
→ **FUZZY LOGIK**

FUZZY LOGIK [7]

viadee®
IT-Unternehmensberatung

Beispiel: Korrelation

Vage Kriterien

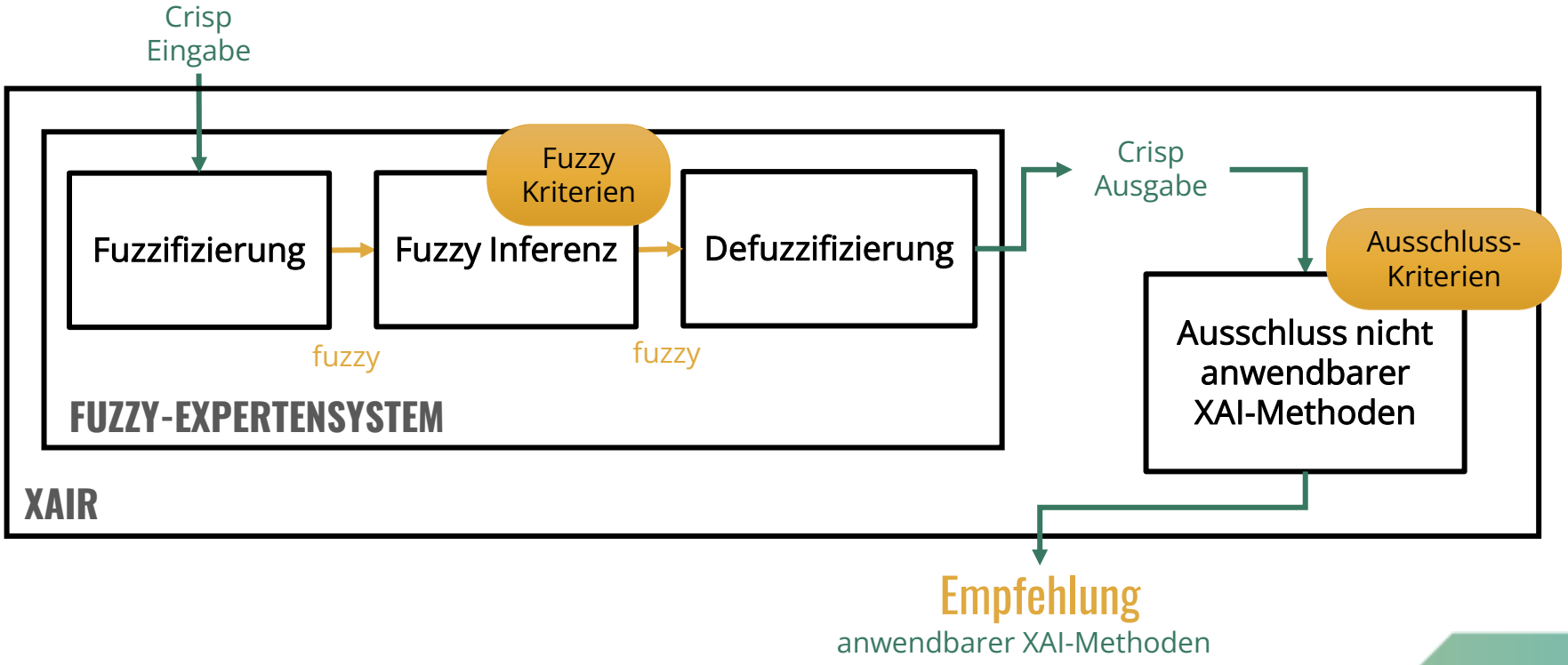


Vage XAI-Methoden-bewertungen

Wenn die Korrelation mittelstark ist, ist Methode X eher nicht geeignet.

→ Regel wird mit einem Gewicht von 0.75 aktiviert

SYSTEMAUFBAU



KONZEPTION / IMPLEMENTIERUNG

FUZZY-EXPERTENSYSTEM

INFERENZ

➡ 38 eignungsreduzierende Regeln (48 insgesamt)

Kriterium	Kriteriums-wert	Methode					
		PDP+ ICE	ALE	PFI	SHAP	Anchors	CF Proto
Korrelation	L	VH	L	-	-	-	-
	M	-	-	L	L	L % 0.6	L % 0.6
	H	L	VH	VL	VL	L	L
Korrelation der FOI	L	VH	L	-	-	-	-
	M	L	H	-	-	-	-
	H	VL	VH	L	L	L % 0.6	L % 0.6
Diskretisierbarkeit	L	-	L % 0.6	-	-	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	-	-
Locale Erklärung	0	-	-	-	-	-	-
	1	VH	-	-	VH	VH	VH

foi_available[True] AND corr_foi[L] -> [PDP + ICE[VH], ALE[L]]
prep_time[L] -> [PFI[L], SHAP[L], Anchors[VL], CFProto[VL]]
foi_available[True] AND corr_foi[H]) AND discr_foi[H] -> [ALE[VH]]
discr[M] -> [Anchors[L]]
corr[L] AND discr[L] -> [ALE[VL]@0.70%]
foi_available[True] AND corr_foi[L]) AND discr_foi[L] -> [ALE[VL]]
perf_pref[H] AND dur_call[H] -> [PFI[L], SHAP[L], Anchors[VL], CFProto[V
foi_available[True] AND discr_foi[L] -> [ALE[L], Anchors[VL]]
init_bb[False] -> [PDP + ICE[M], ALE[M], PFI[M], SHAP[M], Anchors[M], CF
foi_available[True] AND corr_foi[H]) AND discr_foi[M] -> [ALE[H]]
corr[H] -> [PDP + ICE[L], ALE[VH], PFI[VL], SHAP[VL], Anchors[L], CFProto
perf_pref[H] AND dur_call[M] -> [Anchors[L], CFProto[VL]]
init[True] -> [PDP + ICE[M], ALE[M], PFI[M], SHAP[M], Anchors[M], CFPro NOOP[M]]
perf_pref[M] AND dur_call[H] -> [PFI[L]@0.50%, SHAP[L]@0.50%, Anchors[VL]@0.50%, CFProto[VL]@0.50%]
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[L]@0.60%]
discr[L] -> [ALE[L]@0.60%, Anchors[VL]]
ordinal_feat[True] AND discr[M] -> [CFProto[L]@0.60%]
corr[L] -> [PDP + ICE[VH], ALE[L]]
corr[H] AND discr[H] -> [ALE[VH]@0.70%]
scope_global[True] -> [PDP + ICE[VH], ALE[VH], PFI[VH], SHAP[VH]]
foi_available[True] AND discr_foi[M] -> [Anchors[L]]
corr[L] AND discr[H] -> [ALE[L]@0.50%]
discr[H] -> [NOOP[M]]
scope_local[True] -> [PDP + ICE[VH], SHAP[VH], Anchors[VH], CFProto[V
ordinal_feat[True] AND corr[VL] -> [CFProto[VL]@0.60%]
corr[M] AND discr[H] -> [ALE[H]@0.70%]
perf_pref[M] AND num_feat[M] -> [Anchors[L]@0.50%]
foi_available[True] AND corr_foi[M] -> [PDP + ICE[L], ALE[H]]
ordinal_feat[True] AND discr[L] -> [CFProto[VL]@0.80%]
perf_pref[H] AND num_feat[M] -> [Anchors[L]]
corr[H] AND discr[M] -> [ALE[H]@0.70%]
corr[M] -> [PFI[L], SHAP[L], Anchors[L]@0.60%, CFProto[L]@0.60%]
perf_pref[M] AND num_feat[H] -> [PFI[L]@0.50%, SHAP[VL]@0.50%, Anchors[VL]@0.50%, CFProto[L]@0.50%]
foi_available[True] AND corr_foi[H] -> [PDP + ICE[V
Anchors[L]@0.60%, CFProto[L]@0.60%]

FUZZY-EXPERTENSYSTEM

MAX-MIN-INFERENZ & DEFUZZIFIZIERUNG

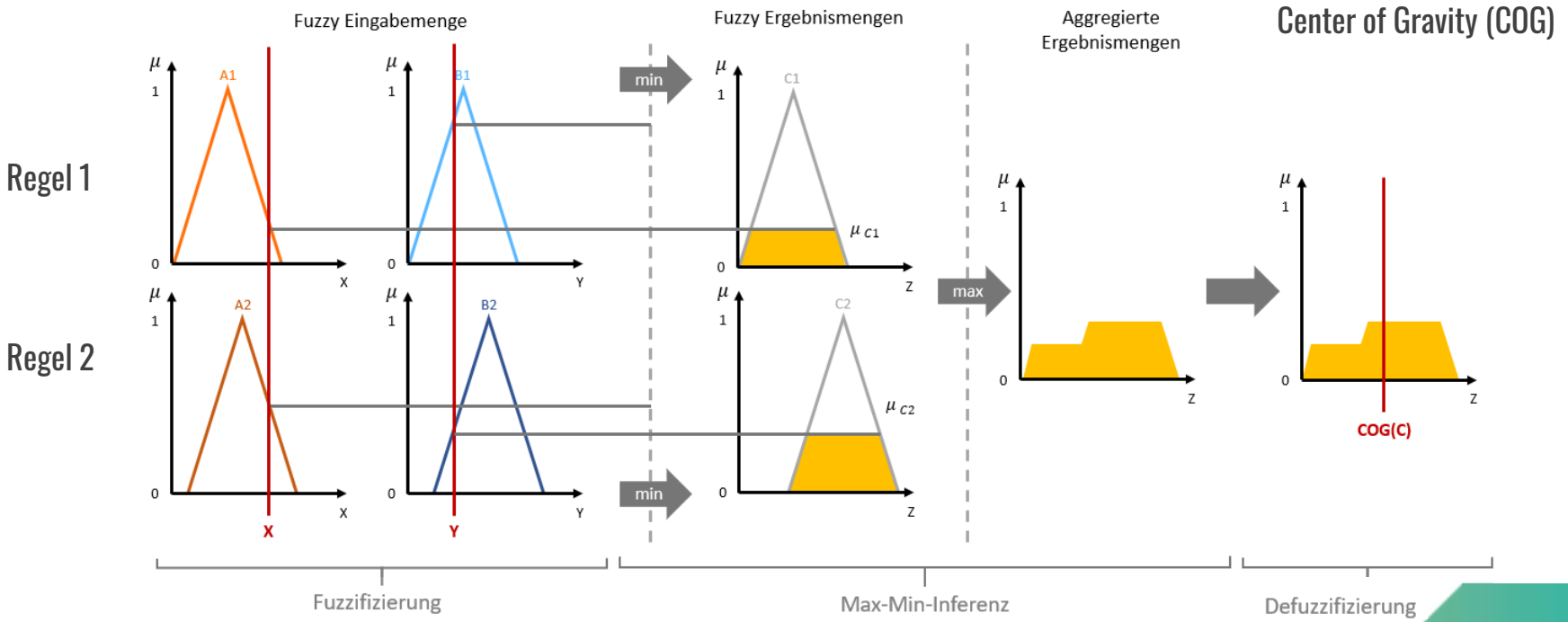


Abbildung angelehnt an [8]

BOOLESCHES AUSSCHLUSSSYSTEM

viadee®
IT-Unternehmensberatung

- Methodenanwendung möglich:
  Eingabevariable >= Methodenbewertung

Ausschlusskriterium	Methode					
	PDP+ ICE	ALE	PFI	SHAP	Anchors	CF Proto
Verfügbarkeit des Modells	0	1	1	0	0	0
Klassifikationsaufgabe	0	0	0	0	1	1
Erhalt der Klassen-wahrscheinlichkeiten	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

IMPLEMENTIERUNG

BACKEND

Python

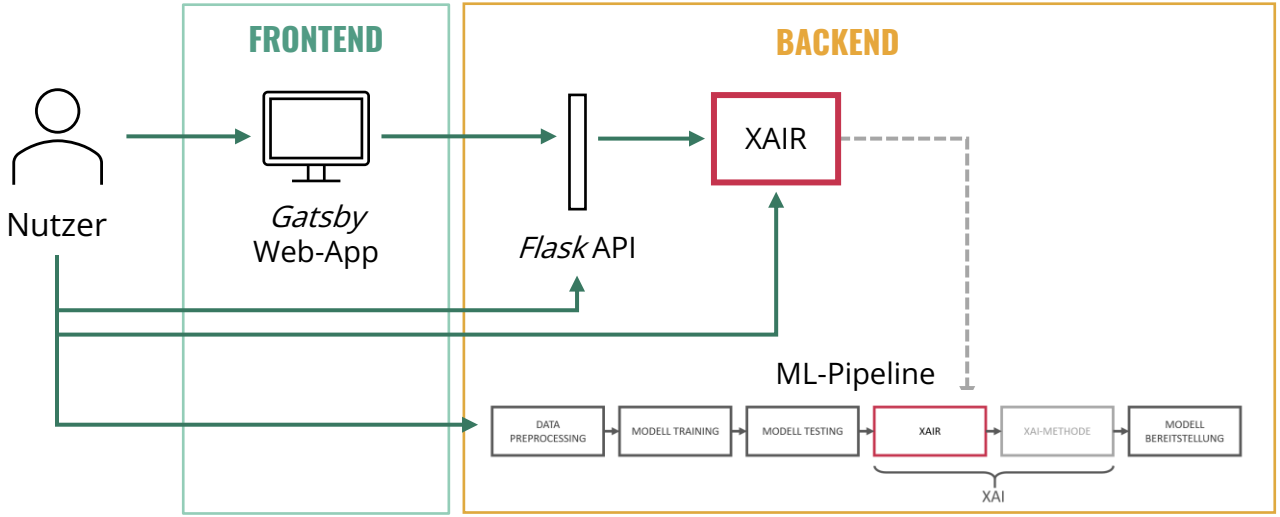
Flask (API)

scikit-fuzzy

FRONTEND

React mit GatsbyJS

(TypeScript)



viadee®
IT-Unternehmensberatung

ERGEBNIS DES XAIRS

Verfügbar unter

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

EVALUATION
NUTZERBEFRAGUNG

viadee®
IT-Unternehmensberatung

- ✓ Übersichtliche Darstellung des Empfehlungsergebnisses
- ✓ Nachvollziehbare Begründung der Empfehlungsentscheidung anhand der Eingaben
  - → Ergebnisse vergleichbar
  - → Relevante Aspekte der Auswahl aufzeigen
- ✓ Benutzerfreundlichkeit
  - → Intuitiv
  - → Ohne tiefe XAI-/ML-Kenntnisse bedienbar
- ✓ Förderung der Bereitschaft der Anwendung von XAI
  - → Zeitersparnis bei Informationsbeschaffung

LIMITATIONEN & HERAUSFORDERUNGEN

viadee®
IT-Unternehmensberatung

- Eingabe benötigter Parameter
  - Schwierigkeiten der fuzzy Einschätzung
  - Minimierung zusätzliche Verzerrungen

  → Umsetzung der automatisierten Datenanalyse
- Beachtung weiterer HCI-, UI- und UX-Aspekte
- Erweiterung der Evaluation auf
  - Qualität der resultierenden Erklärung
  - Umsetzbarkeit der Empfehlung

- Umsetzung nicht implementierter Anforderungen
- Erweiterung des Prototypen ….

  … Feel free to join: https://github.com/viadee/xair

EINORDNUNG IN FORSCHUNGSSTAND

viadee®
IT-Unternehmensberatung

- Ermittlung der Eignung einer XAI-Methode
- Identifikation eignungsbeeinflussender Kriterien
- Erstes Expertensystem im Kontext XAI
  - Organisation existierenden Wissens ausgewählter XAI-Methoden
  - Generalisiert anwendbar
  - Dynamisch erweiterbar (XAI-Methoden, Kriterien)
  - Bietet begründete Empfehlung geeigneter XAI-Methoden

XAIR: „Treibmittel“ zur Auseinandersetzung mit XAI

viadee®
IT-Unternehmensberatung

VIELEN DANK FÜR IHRE AUFMERKSAMKEIT

viadee®
IT-Unternehmensberatung

REFERENZEN

[1] 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-automationinsight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-womenidUSKCN1MK08G. (Zugriff am: 14.10.2020).

[2] Belle, V. & Papantonis, I. (2020), 'Principles and Practice of Explainable Machine Learning', arXiv preprint arXiv:2009.11698 .

[3] Vilone, G. & Longo, L. (2020), 'Explainable Artificial Intelligence: A Systematic Review', arXiv preprint arXiv:2006.00093 .

[4] 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.

[5] 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.

[6] 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'.

[7] Miller, T. (2019), 'Explanation in Artificial Intelligence: Insights from the Social Sciences', Artificial intelligence 267, 1–38.

[8] Zadeh, L. A. (1975), 'Fuzzy logic and approximate reasoning', Synthese 30 (3) pp. 407–428.

[9] 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.