Konzeption und Implementierung eines Empfehlungssystems für die Auswahl und Anwendung von XAI-Methoden

Verena Barth

Hochschule der Medien, Nobelstraße 10, 70569 Stuttgart

Zusammenfassung. 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.

Schlüsselwörter: AI · XAI · XPS · Fuzzy XPS · Recommendation System

1 Einleitung

1.1 Problemstellung

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. [16]
Dieses automatisierte, maschinelle Lernen komplexer Strukturen und Beziehungen aus sehr großen Datensätzen ist sehr leistungsfähig und gilt als Schlüsseltechnologie für kognitive Systeme [10].

Um die gewünschte Genauigkeit der Vorhersagen zu erreichen, werden statt menschenverständlicher Entscheidungssysteme häufig tiefe neuronale Netze 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 beim 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. [3]

Explainable AI (XAI) 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. In der jüngsten Forschung wurde eine Vielzahl von XAI-Methoden und -Implementierungen in Form von eigenständigen Prototyp-Lösungen vorgestellt [37]. Jedoch werden diese in der Praxis kaum angewendet. Dies liegt zum einen daran, dass XAI ein neuer, sich rasant entwickelnder Bereich ist, und zum anderen, dass das bereits existierende Wissen verstreut ist und organisiert werden muss [37].

Einige wissenschaftliche Publikationen klassifizieren zwar diverse Methoden, liefern allerdings selten konkrete Richtlinien oder Ratschläge für ihre Anwendung. Staatliche Behörden, wie bspw. [11], weisen auf die Relevanz der Erklärbarkeit von KI-Systemen hin und darauf, dass bei der Auswahl der methodenspezifischen Eigenschaften der Eingaben berücksichtigt, und plausible Erklärungen ermöglicht werden müssen. Es werden allerdings keine Informationen geboten, die bei einer erfolgreichen Auswahl und Anwendung einer in diesem Sinne geeigneten XAI-Methode helfen. 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 [1].

1.2 Ziele

Zur Förderung des Einsatzes von XAI wird 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 der Forschungsfrage nachzugehen, wird ein XAI-Empfehlungssystem implementiert, das nachfolgend XAIR (XAI Recommender) genannt wird. Für dessen Erstellung 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?

Der XAIR gibt dem Nutzer für ein entwickeltes ML-Modell nachvollziehbar begründete Empfehlungen geeigneter XAI-Methoden. Bei dem Modell, das Klassifikations- oder Regressionsaufgaben anhand tabellarischer Daten löst, kann es sich dabei um eine Black-Box handeln.
Die Zielgruppe des XAIRs besteht aus den Modellentwickelnden, 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.

1.3 Methodik

Um dem gestaltenden, explorativen Charakter der Forschungsfrage gerecht zu werden, wird für die Erstellung des XAIR der lösungsorientierte Design Science Research (DSR) Ansatz in Erwägung gezogen, der durch [17] 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. Da DSR Defizite bei der methodischen Unterstützung der Problemformulierung und -darstellung hat, orientiert sich diese Arbeit an dem Requirements Driven DSR Framework von [7]. Der DSR Ansatz wird dabei 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.

Um einen Überblick über den Bereich XAI, XAI-Methoden und Empfehlungssysteme zu erhalten, wird zudem eine Literaturrecherche nach [38] durchgeführt. Dafür wird mittels einer Stichwortsuche zuerst relevante Literatur mit einem Fokus auf möglichst aktuelle und wissenschaftliche Artikel identifiziert. Ausgehend von [27], was einen guten Überblick über XAI-relevante Terminologien und populäre XAI-Methoden bietet, wird eine Rückwärts- bzw. Vorwärtssuche [38] angewandt.

Die Anfordungsanalyse, sowie die finale Evaluation des Prototyps findet in halbstrukturierten Online-Einzelinterviews statt. Diese offene, qualitative Interviewmethode eignet sich für problemzentrierte Befragungen, zur Prüfung bestehender Hypothesen und zur Gewinnung weiterer Einsichten [8].
Es nehmen jeweils die gleichen fünf Personen mit unterschiedlichen Kenntnissen der Bereiche ML und XAI teil, welche die Zielgruppen des Empfehlungssystem repräsentieren.

2 Verwandte Arbeiten

Im Kontext der Fragestellung relevante Arbeiten bestehen aus wissenschaftlichen Zusammenfassungen des Forschungsbereichs XAI, Richtlinien zur Förderung von Interpretierbarkeit in KI-Systemen und kommerziellen Tools, die XAI automatisiert anwenden. Dabei wird berücksichtigt, dass die für den XAIR identifizierte Zielgruppe keine tieferen ML-Kenntnisse aufweist und, dass die Arbeiten Auskunft über modellagnostische und daten- und domänenunabhängige XAI-Verfahren geben.

Wissenschaftliche Literaturzusammenfassungen, bspw. [27,15], schaffen einen Überblick existierender XAI-Methoden, bieten allerdings keine Orientierung bei der Auswahl einer Passenden. Das Lesen und Verstehen setzt zudem ML- und datenwissenschaftliche Fachkenntnisse voraus.
Obwohl Behörden, bspw. [11], und Unternehmen auf interpretierbare ML-Modelle drängen, geben sie keine umsetzbaren Richtlinien oder konkrete Vorschläge zur Verwirklichung vor [1]. Andere Richtlinien [23,19] 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. Ein statischer Orientierungsbaum [21] bietet eine gute Unterstützung bei der ersten Auseinandersetzung mit XAI-Methoden, schließt aber lediglich unbegründet Methoden aus der Empfehlungsmenge aus und bietet ebenfalls keine Hilfe bei ihrer Anwendung.

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 eignungs-

beeinflussende Kriterien ermittelt, hinsichtlich welcher 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 Auswahl und Anwendung einer XAI-Methode berücksichtigt. Kommerzielle Tools, wie bspw. [12,14,9,35,34], 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.

3 Konzeption

Um dem Nutzer geeignete XAI-Methoden vorschlagen zu können, muss zunächst geklärt werden, was unter der Eignung einer XAI-Methode verstanden wird und anhand welcher Kriterien man diese beurteilen kann.

3.1 Beurteilung der Eignung einer XAI-Methode

Die Eignung einer XAI-Methode lässt sich nicht aufgrund der Qualität der resultierenden Erklärung beurteilen, da Erklärungen kontextbezogen und durch soziale Überzeugungen und kognitive Verzerrungen sehr subjektiv sind [25].
Daher wird sie anhand von Eigenschaften des Modell-, Daten- und Nutzungskontexts definiert, 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.

3.2 Identifikation von Kriterien zur Beurteilung der Eignung von XAI-Methoden

Die eignungsbeeinflussenden Kriterien, die dem XAIR als Eingabeparameter dienen, werden in einer Literaturrecherche und durch eine initiale Anforderungsanalyse ermittelt. Sie lassen sich unterteilen in Ausschlusskriterien, deren binäre Erfüllung für die Anwendung einer XAI-Methode obligatorisch ist, und eignungsbeeinflussende Kriterien, welche sich auf die Eignung einer XAI-Methode einzeln oder in Kombination mit anderen auswirken.

Konkrete Kriterien werden aus den XAI-Methoden abgeleitet, die in den XAIR integriert sind. Zu beachten ist, dass diese implementierungsspezifisch sind, aber nun als repräsentativ für die XAI-Methode angesehen werden. Für die initiale Methodenauswahl werden die in den Richtlinien von [19] und [4] zur Anwendung empfohlen Verfahrensarten, sowie die laut [5] häufig im Deployment eingesetzten *Counterfactual Explanations* berücksichtigt.

Als Visualisierungstechniken sind *Partial Dependence Plot (PDP)* in Kombination mit *Individual Conditional Expectation (ICE)* [32], und *Accumulated Local Effects (ALE)* [18] vertreten. *Anchors* und *Counterfactuals guided by Prototypes (CFProto)* von [33] werden als lokale Methoden, und *Permutation Feature Importance (PFI)* von [20,29] als globales Verfahren hinzugefügt. Die Auswahl wird um *SHapley Additive exPlanations (SHAP)* von [33] ergänzt, die sowohl lokale, als auch globale Erklärungen liefern.

Die anhand dieser XAI-Methoden identifizierten Ausschlusskriterien belaufen sich auf:

- Verfügbarkeit des Modells [1]
- Klassifikationsaufgabe [2]
- Erhalt der Klassenwahrscheinlichkeiten
- Zugriff auf Labels
- Zugriff auf Preprocessing Operationen

Bei den eignungsbeeinflussenden Kriterien des Datenkontexts wird unterschieden, ob sie sich auf den Gesamtdatensatz, oder nur auf kritische Features beziehen. Diese werden nachfolgend FOI („Features of Interest“) genannt. Sie können eine menschliche Voreingenommenheit widerspiegeln, die sich negativ auf die Modellentscheidung auswirken kann (z.B. bzgl. der ethnischen Herkunft, des Geschlechts oder der Religion [2]). Dadurch bieten sie das Potential zur Diskriminierung und bedürfen besonderer Aufmerksamkeit.

Die besondere Berücksichtigung dieser Features und ihr ausführliches Examinieren durch bspw. Visualisierungstechniken ist zur Vermeidung diskriminierender Modelle sinnvoll.

Die identifizierten, eignungsbeeinflussenden Eigenschaften des Modell-, Daten- und Nutzungskontexts sind nachfolgend aufgelistet:

- Korrelation
- Korrelation der FOI
- Diskretisierbarkeit [3]
- Diskretisierbarkeit der FOI [3]
- Anzahl der Features
- Zugriffszeit des Modells/der Vorhersagefunktion
- Vorhandensein ordinaler Features
- Präferenzen der Performance
- Präferenz einer globalen Erklärung
- Präferenz einer lokalen Erklärung
- Vorbereitungsaufwand

[1] Ist kein Zugriff auf das Modell gegeben, muss die Vorhersagefunktion verfügbar sein.

[2] Unter der Annahme, dass es sich entweder um eine Klassifikations- oder eine Regressionsaufgabe handelt.

[3] Diskretisierbarkeit ist gegeben, wenn die Datenpunkte einer Feature-Verteilung in gleichgroße Intervalle (EqualWidth Binning) mit ähnlicher Anzahl der darin liegenden Datenpunkte, oder in bspw. Dezile (EqualFrequency Binning) mit ähnlicher Breite aufgeteilt werden können.

3.3 Architektur des XAIRs

Strukturierung aufgrund der verwendeten Logik Der XAIR operiert in zwei Stufen. Zunächst bestimmt er mithilfe eines Expertensystems [30] durch die Verwendung von fuzzy [28] Regeln die Methodeneignung anhand der eignungsreduzierenden Kriterien. Die darin angewandte Fuzzy-Logik [39] adressiert die diversen Probleme der Ungenauigkeit, die im Kontext von XAI auftreten: Sie vermeidet die Definition von Schwellwerten der eignungsbeeinflussenden Kriterien (z.B. bei der Korrelation des gesamten Datensatzes), welche für eine exakte Quantifizierung der Eingaben notwendig wären. Sie ermöglicht eine ungefähre Einschätzung subjektiver Kriterien (z.B. dem Vorbereitungsaufwand), oder von Kriterien, deren Werte aufgrund ihrer Unterschiedlichkeit oder einem Mangel an Referenzen schwer möglich ist (z.B. Anzahl der Features). 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 Empfehlungsresultate [39].

Aus der Menge aller vom Fuzzy-Expertensystem zurückgegebenen Methodeneignungen eliminiert der XAIR im zweiten Schritt die XAI-Methoden, welche durch die binären Ausschlusskriterien nicht anwendbar sind.

Aufbau hinsichtlich diverser Nutzungsszenarien Die Konsultation des XAIRs kann manuell über eine Webanwendung erfolgen, wobei die Eingabeparameter über eine grafische Weboberfläche vom Nutzer entgegengenommen werden. Die datenbezogenen Parameter, die Korrelation (der FOI) und Diskretisierbarkeit (der FOI), können nach einer manuell durchgeführten Datenanalyse entweder crisp, oder nach eigener Einschätzung fuzzy sein.

Durch eine Implementierung in Python [31] und eine Containerisierung mit Docker [24] ist der XAIR als Microservice in einer beliebigen Umgebung einsetzbar. Somit kann er minimal-invasiv in eine Komponente einer Kubeflow ML-Pipeline [22] integriert und dort automatisiert verwendet werden. Das ist gewünscht, da durch die Automatisierung der Schritt der Auswahl einer XAI-Methode obligatorisch ist, und die Modellverantwortlichen zu einer Auseinandersetzung mit XAI bewogen werden.
Im Hinblick auf die Verwendung innerhalb einer ML-Pipeline ist eine automatisierte Ermittlung der für eine Ergebnisgenerierung benötigten Dateneigenschaften notwendig. Alle nicht datenbezogenen Eingabeparameter müssen dafür in der Pipelinekonfiguration hinterlegt sein, da sie nicht in einer Datenanalyse bestimmt werden können.

3.4 Konfiguration des Fuzzy-Expertensystems

Identifikation der Ein- und Ausgabevariablen Die Eingabevariablen des Fuzzy-Expertensystems sind die eignungsbeeinflussenden Parameter. Die Ausgaben sind die für den XAIR ausgewählten XAI-Methoden, deren resultierende

Werte die Eignung der einzelnen XAI-Methoden widerspiegeln.
Die Modellierung der Zugehörigkeitsfunktionen der Variablen und die Beurteilung der Mitgliedsgradwerte erfolgt aufgrund von Erfahrungen, persönlichen Einschätzungen und sprachlichen Gewohnheiten nach sachinhaltlichen Gegebenheiten [6]. Die Wertebereiche der Ausgabevariablen, d.h. der Methodeneignungen, werden bspw. in fünf Zugehörigkeitsfunktionen unterteilt: Ungeeignet („VL“), eher ungeeignet („L“), neutral („M“), geeignet („H“) und sehr gut geeignet („VH“).

Erstellung der Regelbasis 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.
Sie besteht aus 38 Fuzzy-Regeln, die mindestens die Eignung einer XAI-Methode beeinflussen.

Bei der Formulierung dieser Regeln wurde berücksichtigt, dass alle Methoden initial mit „M“ bewertet werden, um sicherzustellen, dass jeder Ausgabeparameter des Fuzzy-Systems einen definierten Zustand hat.
Bei der Interpretation der Empfehlungsergebnisse ist zu beachten, dass ein resultierender, crisp Empfehlungswert von 5 (fuzzy „M“) durch einen Ausgleich positiver und negativer Bewertungen oder durch eine Abwesenheit von eignungsreduzierenden Kriterienbewertungen entstehen kann. Er sagt folglich nicht zwangsläufig eine mittelmäßige, sondern eher eine neutrale Eignung aus und sollte daher nicht als schlecht interpretiert werden. Das Ergebnis der Eignung einer XAI-Methode ist nicht absolut, sondern relativ zur Eignung der anderen zu sehen.

Um eine Empfehlung nicht anwendbarer XAI-Methoden zu vermeiden, werden die Methoden auch hinsichtlich der booleschen Ausschlusskriterien bewertet. Dieses Wissen ist nicht innerhalb der Fuzzy-Wissensbasis gespeichert.

Wahl der Defuzzifizierungsstrategie Um einen konkreten, crisp Eignungswert aus den fuzzy Ergebnismengen der aktivierten Regeln zu erhalten, wird die Center of Gravity (COG) Defuzzifizierungsmethode angewendet. Alternative Maxima-Methoden, die häufig in wissensbasierten Fuzzy-Systemen Anwendung finden [36], ignorieren alle nicht maximal aktivierten Regeln. Da alle durch die Kriterien aktivierten Regeln für in die Eignungsbewertung der XAI-Methode berücksichtigt werden sollen, wird für die Defuzzifizierung des Empfehlungsergebnisses daher die COG Strategie gewählt.

4 Resultate

4.1 Evaluation des Prototyps

Um den Nutzen und die Ergebnisqualität für den Nutzer und die organisationsbezogene Produktivität zu beurteilen (vgl. Bewertungskriterien für Expertensysteme nach [26]), wird der webbasierte XAIR nach dem Ansatz des Requirements Driven DSR anhand der Erfüllung der Anforderungen beurteilt. Nach einer selbstständigen, zeitlich unbegrenzten Nutzung des XAIRs mit einem konkreten

Anwendungsfall finden dafür Nutzerbefragungen in Form von halbstrukturierten Online-Einzelinterviews statt. Der Fokus lag dabei auf der Beantwortung der Forschungsfrage und der Erfüllung der Anforderungen der Benutzerfreundlichkeit, der Übersichtlichkeit der Empfehlungsergebnisse und ihrer Nachvollziehbarkeit, sowie der Förderung der Bereitschaft der Anwendung von XAI.

Nutzen Der Aufbau des XAIR wurde generell als intuitiv, nutzerfreundlich, optisch ansprechend und als „leichtgewichtig in der Anwendung“ wahrgenommen. Durch textuelle Hilfestellungen sind für seine Verwendung keine tieferen ML-/XAI-Kenntnisse notwendig. Die empfehlungsunabhängige Möglichkeit, ohne Zuhilfenahme (aber mit Angabe) weiterer Quellen tiefere Methodeneinblicke zu erhalten, wurde sehr wertgeschätzt.

Ergebnisqualität Die Ergebnisdarstellung, sowie die Begründung der Empfehlungsentscheidung und der Eignungsvergleich der XAI-Methoden anhand der Kriterien, wurden von allen Teilnehmenden als sehr positiv und übersichtlich empfunden. Die Empfehlungen sind für alle nachvollziehbar.
Obwohl sich keine Person im Detail mit den zusätzlichen Methoden- und Implementierungsinformationen auseinandergesetzt hat, wurden ihre Bereitstellung sehr positiv wahrgenommen. Sie macht „den Eindruck, als dass es sich lohnt, diese Info durchzulesen“, da sie „sehr gezielt“ ist und durch die trichterförmige Aufbereitung in kurze Abschnitte „recht bekömmlich“ wirkt.

Produktivität Die Zeitersparnis bei der Methodenauswahl wurde besonders von ML-erfahrenen Personen geschätzt. Eine dieser „weiß, was es bedeutet, sich so Wissen erstmal zu erarbeiten und deswegen kann [sie] ganz gut einschätzen, wieviel Zeit [...] das hier spart“. Nach eigener Aussage „extrem viel Zeit“.
Bei der vagen Eingabe der datenspezifischen Eigenschaften hatten drei Nutzer (unabhängig ihres Expertenwissens) durch Fehlen eines „Ankerpunktes“ bzw. eines Auswahlbereichs Probleme. Eine weitere empfand die fuzzy Angabe als kritisch, da dadurch eine Voreingenommenheit des Nutzers bezüglich der Daten für die Auswahl einer XAI-Methode berücksichtigt wird.

4.2 Diskussion der Ergebnisse

Grundsätzlich wurden die Erwartungen aller Teilnehmenden an das XAI-Empfehlungssystem maßgeblich erfüllt und die einer Person, durch die Bereitstellung von strukturiertem, tieferen Methodenwissen und Implementierungsvorschlägen und -hinweisen, sogar übertroffen. Alle würden den XAIR nochmals bzw. mehrmals nutzen. Eine mit XAI vertraute Person 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 [des Modells] ist doch schon bei 95%, was muss ich dann noch erklären.“

Zukünftig kann das durch die Evaluation identifizierte Problem fehlender Unterstützung bei der subjektiven Einschätzung der datenbezogenen Kriterien durch die Umsetzung eines bereits entwickelten Konzepts der automatisierten Datenanalyse adressiert werden.

Auch eine Ausweitung und Verlängerung der Nutzerstudie wäre denkbar, um die Anwendbarkeit der XAI-Empfehlung mit der gegebenen Information und die situationsspezifische Qualität der resultierenden Erklärung zu evaluieren. Diese war zum einen aus Zeitgründen nicht umsetzbar und zum anderen, da die Quantifizierung des Verständlichkeitsgrads einer Erklärung für den Menschen noch nicht (mathematisch) formalisierbar und messbar ist [13].

Die Evaluation deutet an, dass der resultierende XAIR die erhobenen Anforderungen und Systemziele erfüllt. Zusammenfassend unterstützt er den Nutzer durch nachvollziehbar erklärte und untereinander vergleichbare Empfehlungen, die ihn bzgl. der für eine Methodenauswahl zu berücksichtigenden Kriterien sensibilisieren. Die Aufbereitung tieferen Wissens spart ihm Zeit bei der weiteren Informationsbeschaffung, welche für eine Methodenanwendung notwendig ist.

5 Fazit

Es 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) konzipiert und implementiert, welches dem Nutzer ohne tiefe ML-Kenntnisse für den Anwendungskontext geeignete XAI-Methoden vorschlägt.
Der Source Code ist unter der folgenden URL öffentlich verfügbar:
`https://github.com/viadee/xair`

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 bezüglich XAI reduziert. Die Evaluation weist darauf hin, dass der Nutzer durch die Zeitersparnis dieser Empfehlungen auch zur erneuten Konsultation des XAIRs und zur anschließenden Anwendung der XAI-Methode motiviert wird.

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

Der resultierende, prototypische XAIR liefert schnell nutzungskontextspezifische Empfehlungen geeigneter XAI-Methoden. 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 bezüglich der XAI Informationsbeschaffung, wird der Nutzer erfolgreich bei der Auswahl einer XAI-Methode unterstützt. Außerdem wird er bzgl. der Eigenschaften des Nutzungskontexts sensibilisiert, die es bei der Auswahl zu beachten gilt.
Durch seine umfassenden und nachvollziehbaren Empfehlungen, seine nutzerfreundliche Anwendung und seine Erweiterbarkeit, hebt er sich im Besonderen von aktuellen Alternativen wie dem statischen Orientierungsbaum von [21] ab, der unbegründet Empfehlungen für XAI-Methoden auf Basis weniger Kriterien liefert.

Zusammenfassend kann gesagt werden, dass das konzipierte XAI-Empfehlungssystem das Problem der fehlenden Unterstützung bei der Auswahl und Operationalisiertung 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.

Literatur

1. AlgorithmWatch: AI Ethics Guidelines Global Inventory (2019), [Online]. Verfügbar unter https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/ (Zugriff am: 23.04.2021)
2. Antidiskriminierungsstelle des Bundes: Allgemeines Gleichbehandlungsgesetz: AGG (2006)
3. 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.: Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion **58**, 82–115 (2020)
4. Belle, V., Papantonis, I.: Principles and Practice of Explainable Machine Learning. arXiv preprint arXiv:2009.11698 (2020)
5. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M., Eckersley, P.: Explainable machine learning in deployment. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 648–657 (2020)
6. Böhme, G.: Fuzzy-Logik: Einführung in die algebraischen und logischen Grundlagen. Fuzzy-Logik, Springer Berlin Heidelberg, Berlin, Heidelberg (1993). https://doi.org/10.1007/978-3-642-86785-9
7. Braun, R., Benedict, M., Wendler, H., Esswein, W.: Proposal for requirements driven design science research. In: Donnellan, B., Helfert, M., Kenneally, J., VanderMeer, D., Rothenberger, M., Winter, R. (eds.) New Horizons in Design Science: Broadening the Research Agenda, vol. 9073, pp. 135–151. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18714-3_9
8. Buber, R., Holzmüller, H.H.: Qualitative Marktforschung: Konzepte - Methoden - Analysen. Gabler-Lehrbuch, 2. Aufl., Gabler, Wiesbaden (2009)
9. DataRobot: Model Interpretability (2020), [Online]. Verfügbar unter https://www.datarobot.com/wiki/interpretability/ (Zugriff am: 12.10.2020)
10. Döbel, I., Leis, M., Molina Vogelsang, M., Neustroev, D., Petzka, H., Riemer, A., Rüping, S., Voss, A., Wegele, M., Welz, J.: Maschinelles Lernen: Eine Analyse zu Kompetenzen, Forschung und Anwendung (2018). https://doi.org/10.1007/978-3-642-60452-2_3
11. Federal Office for Information Security: AI Cloud Service Compliance Criteria Catalogue (AIC4) (2021)
12. Google: Introduction to AI Explanations for AI Platform (2020), [Online]. Verfügbar unter https://cloud.google.com/ai-platform/prediction/docs/ai-explanations/overview (Zugriff am: 12.10.2020)
13. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., Giannotti, F.: A Survey Of Methods For Explaining Black Box Models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
14. H2O.ai: Machine learning interpretability: Interpretability in h2o driverless ai (2020), [Online]. Verfügbar unter https://www.h2o.ai/products-dai-mli/ (Zugriff am: 12.10.2020)
15. Hall, P., Gill, N.: An Introduction to Machine Learning Interpretability. O'Reilly Media, Incorporated (2019)
16. Hamon, R., Junklewitz, H., Sanchez, I.: Robustness and explainability of artificial intelligence. Publications Office of the European Union (2020)
17. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Quarterly (1), 75–105 (2004)

18. Jomar, D.: PyALE. Version 1.0.0.post1, Verfügbar unter https://github.com/DanaJomar/PyALE/blob/master/PyALE/ (2020)
19. Kangur, A.: Explainable ai in practice: How committing to transparency made us deliver better ai products (2020), [Online]. Verfügbar unter https://towardsdatascience.com/explainable-ai-in-practice-6d82b77bf1a7 (Zugriff am: 28.01.2021)
20. Korobov, M., Lopuhin, K.: ELI5. Version 0.10.1, Verfügbar unter https://github.com/eli5-org/eli5 (2019)
21. Kraus, T., Ganschow, L., Eisenträger, M., Wischmann, S.: Erklärbare Künstliche Intelligenz - Anforderungen, Anwendungen, Lösungen. Technologieprogramm KI-Innovationswettbewerb des Bundesministeriums für Wirtschaft und Energie (04 2021)
22. Kubeflow: Overview of Kubeflow Pipelines: Understanding the goals and main concepts of Kubeflow Pipelines (2020), [Online], Verfügbar unter https://www.kubeflow.org/docs/pipelines/overview/pipelines-overview (Zugriff am: 19.10.2020)
23. Leslie, D.: Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. SSRN 3403301 (2019)
24. Merkel, D.: Docker: Lightweight linux containers for consistent development and deployment. Linux J. **2014**(239) (Mar 2014)
25. Miller, T.: Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial intelligence **267**, 1–38 (2019)
26. Miranda, P., Isaias, P., Crisóstomo, M.: 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, pp. 100–109. Springer (07 2011). https://doi.org/10.1007/978-3-642-21672-5_12
27. Molnar, C.: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2020)
28. Nissen, V.: Ausgewählte Grundlagen der Fuzzy Set Theorie. Reihe Ilmenauer Beiträge zur Wirtschaftsinformatik, Arbeitsbericht Nr. 2007-03 (06 2007)
29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011), Version 0.23.2, Verfügbar unter https://github.com/scikit-learn/scikit-learn
30. Puppe, F.: Einführung in Expertensysteme. 2. Aufl., Springer, Berlin, Heidelberg (1991)
31. Python Software Foundation: Python Language Reference. Version 3.9, Verfügbar unter https://www.python.org (2020)
32. SauceCat: PDPbox. Version 0.2.0, Verfügbar unter https://github.com/SauceCat/PDPbox (2018)
33. Seldon: Alibi. Version 0.5.5, Verfügbar unter https://github.com/SeldonIO/alibi (2020)
34. Spinner, T., Schlegel, U., Schafer, H., El-Assady, M.: explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. IEEE transactions on visualization and computer graphics **26**(1), 1064–1074 (2020). https://doi.org/10.1109/TVCG.2019.2934629
35. Uppington, W.: Introducing Truera. [Online], Verfügbar unter https://truera.com/introducing-truera/ (2020), (Zugriff am: 12.10.2020)

36. van Leekwijck, W., Kerre, E.E.: Defuzzification: Criteria and classification. Fuzzy Sets and Systems **108**(2), 159–178 (1999). https://doi.org/10.1016/S0165-0114(97)00337-0
37. Vilone, G., Longo, L.: Explainable Artificial Intelligence: a Systematic Review. arXiv preprint arXiv:2006.00093 (2020)
38. vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., Cleven, A.: Reconstructing the Giant - On the Importance of Rigour in Documenting the Literature Search Process (2009)
39. Zadeh, L.A.: Fuzzy logic and approximate reasoning. Synthese 30 (3) pp. 407–428 (1975). https://doi.org/10.1007/BF00485052