Erklärbare Künstliche Intelligenz - Steigerung der Nachvollziehbarkeit überwachter maschineller Lernverfahren

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

von der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Nadia Burkart (geb. El Bekri)

aus Konstanz

Tag der mündlichen Prüfung:	17.12.2021
Erster Gutachter:	Prof. Dr.-Ing. habil. Jürgen Beyerer
Zweiter Gutachter:	Prof. Dr.-Ing. habil. Marco F. Huber

Abstract

Artificial Intelligence (AI), and in particular the field of machine learning, is becoming increasingly important in our everyday lives. In the future, these methods will increasingly be used, for example, in medicine to help diagnose illnesses or in banking to detect money laundering. Reservations about the use of these methods are very often associated with their high complexity and the associated lack of explainability. Models generated by machine learning methods are considered a "black box" because they are usually incomprehensible to the user. There is a lack of insight into how the model generates the results. The research field of Explainable Artificial Intelligence (XAI) tries to design solutions that make entire models or specific model results comprehensible.

This thesis contributes to the research field of XAI, especially in the subfield of explainable supervised machine learning. The first major contribution of the work is the design of a procedure model which defines different types of explanations that can be generated in this subfield. In total, five types of explanations are distinguished, and these can be roughly divided into model and instance explanations. Another major contribution of the work is its procedures for generating explanations. Based on this procedure model, two procedures were designed: a global surrogate model and a local surrogate model. Starting from a neural network, a global surrogate model is generated in the form of a comprehensible surrogate model. The surrogate is generated using regularization so that it satisfies both the properties of explainability and accuracy. The local surrogate model generates the explainability of a single instance starting from a black box. The goal is to generate an explanation that is close to the decision boundary of the original data instance. This type of explanation allows the user to generate preferred outcomes.

Explanations that make models or specific model results comprehensible are an important step in the process of explainable machine learning. To investigate the comprehensibility of explanations, the utility of generated explanations must be investigated with users. Therefore, another major contribution of the thesis is the investigation of the comprehensibility of explanations in the context of user studies. Different tasks and user groups were included in these investigations.

Kurzfassung

Das Thema Künstliche Intelligenz (KI) und insbesondere das Gebiet der maschinellen Lernverfahren findet immer mehr Einzug in das tägliche Leben. In Zukunft werden die Verfahren verstärkt beispielsweise in der Medizin bei der Diagnose einer Krankheit oder im Bankenwesen beim Aufspüren von Geldwäschetransaktionen unterstützen. Vorbehalte gegenüber dem Einsatz der Verfahren sind oft mit der hohen Komplexität und der einhergehenden fehlenden Nachvollziehbarkeit der Modelle verbunden. Modelle, die durch ein maschinelles Lernverfahren erzeugt werden, gelten als Blackbox, da diese meist für die Anwender nicht nachvollziehbar sind. Es fehlen Erkenntnisse darüber, wie das Modell die Ergebnisse erzeugt. Das Forschungsfeld der Erklärbaren Künstlichen Intelligenz versucht, Lösungen zu konzipieren, die die Nachvollziehbarkeit von gesamten Modellen oder bestimmten Modellergebnissen erhöhen.

Die vorliegende Arbeit leistet einen Beitrag zum Forschungsfeld der Erklärbaren KI – insbesondere im Teilbereich des erklärbaren überwachten maschinellen Lernens. Der erste wesentliche Beitrag der Arbeit umfasst den Entwurf eines Vorgehensmodells. Dieses definiert unterschiedliche Arten von Erklärungen, die im Bereich des überwachten maschinellen Lernens generiert werden können. Insgesamt werden fünf Arten von Erklärungen unterschieden, die sich grob in Modell- und Instanz-Erklärungen einteilen lassen. Ein weiterer wesentlicher Beitrag der Arbeit sind die entworfenen Verfahren zur Erstellung von Erklärungen. Basierend auf diesem Vorgehensmodell wurden zwei Verfahren entworfen: ein globales Surrogat-Modell und ein lokales Surrogat-Modell. Ausgehend von einem neuronalen Netz wird ein globales Surrogat-Modell in Form eines nachvollziehbaren Modells erzeugt, das sich somit den

Modell-Erklärungen zuweisen lässt. Das Surrogat wird mithilfe der Regularisierung generiert, sodass dieses sowohl die Eigenschaften der Nachvollziehbarkeit als auch die der Genauigkeit erfüllt. Das lokale Surrogat-Modell hingegen versucht, die Nachvollziehbarkeit eines einzelnen Ergebnisses ausgehend von einer Blackbox zu erzeugen, und lässt sich den Instanz-Erklärungen zuordnen. Das Ziel ist es, eine Erklärung zu erhalten, die nahe an der Entscheidungsgrenze zur ursprünglichen Dateninstanz liegt. Diese Art der Erklärung ermöglicht es dem Anwender, bevorzugte Ergebnisse des Modells zu erzeugen.

Erklärungen, die Modelle oder bestimmte Modellergebnisse nachvollziehbar gestalten, sind ein wichtiger Schritt im Prozess des erklärbaren maschinellen Lernens. Um die Nachvollziehbarkeit von Erklärungen zu untersuchen, ist es zwingend notwendig, den Nutzen generierter Erklärungen mit Anwendern zu analysieren. Daher umfasst ein weiterer wesentlicher Beitrag der Arbeit die Untersuchung der Nachvollziehbarkeit von Erklärungen im Rahmen von Benutzerstudien. Dabei wurden sowohl unterschiedliche Aufgabenstellungen als auch Anwendergruppen in die Untersuchungen miteinbezogen.

Danksagung

Die vorliegende Arbeit wurde am Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung IOSB in Kooperation mit dem Karlsruher Institut für Technologie (KIT) angefertigt. An dieser Stelle möchte ich mich besonders bei den nachfolgenden Personen bedanken, die mich bei der Anfertigung meiner Dissertation unterstützt haben.

Zunächst möchte ich mich bei Herrn Prof. Dr.-Ing. habil. Jürgen Beyerer, Leiter des Fraunhofer IOSB, bedanken, der mir die Möglichkeit gegeben hat, diese Arbeit unter seiner hervorragenden Leitung und in Kooperation mit dem KIT durchzuführen. Herrn Prof. Dr.-Ing. habil. Marco Huber danke ich vor allem für die Themenstellung, die exzellente Betreuung und seine ausdauernde Diskussions- und Hilfsbereitschaft. Für die Möglichkeit, stets neue Ideen auf Augenhöhe zu erörtern, möchte ich mich herzlich bedanken. Ebenso möchte ich mich bei Frau Dr. Elisabeth Peinsipp-Byma für die Möglichkeit bedanken, diese Arbeit in ihrer Abteilung zu verfassen. Ihre pragmatischen Ratschläge und die wertvollen Anregungen haben wesentlich zum Gelingen dieser Arbeit beigetragen. Auch möchte ich den Studenten, die ich betreut habe, darunter besonders Philipp M. Faller für die sehr gute Arbeit danken.

Meinen Geschwistern und Freunden danke ich für die Rücksichtnahme und die vielen Ermutigungen. Mein ganz besonderer Dank gilt meinen Eltern, die mir meinen bisherigen Lebensweg ermöglicht haben und die mich bei allem unterstützt haben. Dieser Beistand war mir immer wichtig, vor allem im Verlauf dieser Dissertation.

Tief verbunden und dankbar bin ich meinem Ehemann Martin, ohne dessen bewundernswerte Geduld und Verständnis ein solcher Arbeitsumfang in dieser beschwerlichen Zeit niemals hätte gelingen können. Sein moralischer Beistand und menschlicher Halt haben mir Kraft und Mut zur Anfertigung und Vollendung dieser Dissertation gegeben. Gleicher Dank gilt meinem Sohn Maximilian, dessen Lächeln mich durch die schwierige Endphase dieser Arbeit getragen hat. Euch beiden widme ich diese Arbeit.

Karlsruhe, im August 2022 Nadia Burkart

Für Martin und Maximilian

Inhaltsverzeichnis

Notation

Konventionen

$\mathcal{A}, \mathcal{B}, ...$	Mengen (Großbuchstaben), kalligrafisch
$i, j, k, ...$	Ganzzahlige Indexvariablen (Kleinbuchstaben, kursiv)
$X, Y, ...$	Variablen (Grossbuchstaben, kursiv)
$x, y, ...$	Ausprägungen von Variablen (Kleinbuchstaben, kursiv)
α, β	Gewichte von Variablen (Kleinbuchstaben, kursiv)
$\mathbb{N}$	Menge der natürlichen Zahlen inklusive Null
$\mathbb{R}$	Menge der reellen Zahlen

Bezeichnungen

$\mathcal{A}$	Allgemeiner Bezeichner für eine Menge von Regeln
A	Ausprägung einer Menge von Regeln
$\mathcal{B}$	Allgemeiner Bezeichner für eine Menge von Blackbox-Modellen
b	Ausprägung eines Blackbox-Modells
b_θ	Ausprägung eines Blackbox-Modells mit den Parametern θ
γ_{ti}	Zweite Schätzung einer Vorhersage mit Index i und t
C	Allgemeiner Bezeichner für Matrix über das Vorkommen eines bestimmten Musters

$C_{i,j}$	Allgemeiner Bezeichner für Matrix über das Vorkommen eines bestimmten Musters mit Index i für die Instanz und Index j für das Muster
C^T	Transponierte Matrix der Matrix C
D	Allgemeiner Bezeichner für einen Datensatz
d	Distanzmetrik
$\mathcal{E}$	Allgemeiner Bezeichner für eine Menge von Erklärfunktionen
e	Ausprägung einer bestimmten Erklärfunktion
$\mathcal{F}$	Menge häufiger Muster
F	Ausprägung eines häufigen Musters
F_j	Ausprägung eines häufigen Musters mit Index j
$\mathcal{H}$	Menge von Modellen
h	Ausprägung eines Modells
L	Allgemeiner Bezeichner für die Verlustfunktion
L'	Verlustfunktion mit Regularisierungsterm
L''	Verlustfunktion mit Regularisierungsterm durch Proxy-Funktion
λ	Parameter zur Gewichtung u. a. für die Regularisisierungsstärke
μ_{ti}	Erste Schätzung einer Vorhersage mit Index t und i
MAD	Wert der mittleren absoluten Abweichung vom arithmentischen Mittel
Ω	Allgemeiner Bezeichner für die Regularisierung
R^F	Matrix zur Summe aller Differenzen innerhalb eines Musters F
$R^F_{i,j}$	Matrix zur Summe aller Differenzen innerhalb eines Musters F mit Index i und j

$\mathcal{P}$	Wahrscheinlichkeit
ρ	Menge von Bedingungen einer Regel
q_ϕ	Proxy-Netz mit den Parametern ϕ
S	Allgemeiner Bezeichner für ein Fehlermaß
s	Ausprägung eines bestimmten Fehlermaßes
supp(A)	Support einer Regel (auch als Itemset bezeichnet)
Θ	Allgemeiner Bezeichner für eine Menge von Parametern eines Modells
θ	Ausprägung bestimmter Parameter eines Modells
w	Ausprägung eines (Whitebox-)Modells
w_A	Ausprägung eines (Whitebox-)Modells als Regelliste
$\tilde{x}_i$	Median
V	Matrix der Differenz zwischen zwei Ausgaben

Abkürzungsverzeichnis

AI	Artificial Intelligence
AUC	Area Under The Curve
BRL	Bayesian Rule List
DARPA	Defense Advanced Research Projects Agency
DSGVO	Datenschutz-Grundverordnung
KI	Künstliche Intelligenz
KNN	Künstliche Neuronale Netze
LIME	Local Interpretable Model-Agnostic Explanations
MAD	Mean Absolute Deviation
MCR	Model Class Reliance
ML	Maschinelle Lernverfahren
MLP	Multilayer-Perzeptron
OWL	Web Ontology Language

PDP	Partial Dependence Plot
ReLu	Rectified Linear Unit
RF	Random Forest
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
ROC	Receiver Operating Characteristic
SBRL	Scalable Bayesian Rule List
W3C	World Wide Web Consortium
XAI	Explainable Artificial Intelligence

1 Einleitung

Maschinelle Lernverfahren (ML) bilden die Grundlagen für eine Vielzahl von Systemen im Bereich der KI. Die zugrundeliegenden Verfahren im Bereich des maschinellen Lernens konstruieren zumeist komplexe Modelle, die zur Lösung spezieller Aufgaben eingesetzt werden. Im Februar 2020 präsentierte Microsoft das bislang größte neuronale Netz *Turing NLG* zur Sprachgenerierung mit rund 17 Milliarden Parametern [Ras20]. Aufgrund der hohen Komplexität werden die Modelle oft auch als „Blackbox" bezeichnet. In diesem Zusammenhang wird eine Blackbox als ein nicht nachvollziehbares Modell für den Menschen verstanden. KI wird vor allem als Technologie eingesetzt, um in kurzer Zeit einen deutlichen Mehrwert zu erzeugen. Ergebnisse einer Studie zeigen, dass sich der überwiegende Teil der KI-Projekte bereits in weniger als zwei Jahren amortisiert – bei deutschen Unternehmen gar noch etwas schneller als im globalen Vergleich [Del20]. Ein prominentes Beispiel eines KI-Systems ist *AlphaGo Zero* [Sil17]. Das System gewann gegen die weltweit führenden Go-Spieler. Das Spiel, bei dem abwechselnd weiße und schwarze Steine auf das Spielbrett gelegt werden, wurde lange Zeit als zu komplex für ein KI-System angesehen, um die Spielzüge zu erlernen.

KI-Systeme werden ebenso im Gesundheits- oder Finanzwesen – meist zur Entscheidungsunterstützung – eingesetzt. So kann beispielsweise das KI-System *Watson* von IBM Ärzte bei der Diagnose und individuellen Therapie von Krebspatienten unterstützen. Vorbehalte gegenüber den Systemen bleiben bestehen, da diese als Blackbox und damit als nicht nachvollziehbar gelten. Holzinger et al. [Hol19] stellen fest, dass der medizinische Bereich zu den größten Herausforderungen der KI gehört. Für KI-Anwendungen im medizinischen Bereich, die ein tiefes Verständnis verlangen, besteht die Notwendigkeit, nachvollziehbare Entscheidungen zu erzeugen. Eine Diagnose

oder eine Empfehlung für eine Therapie muss notwendigerweise für die Anwender[1] der Systeme, in diesem Fall die behandelnden Ärzte, nachvollziehbar gestaltet sein. Neben der Genauigkeit der Modelle ist das detaillierte Verständnis des Modells und der Ergebnisse essenziell. Holzinger et al. [Hol20] stellen heraus: „Obwohl Menschen dazu neigen, mehr Fehler zu machen, ist die menschliche Intelligenz als solche in der Regel zuverlässiger und robuster gegen katastrophale Fehler, während die KI anfällig für bereits kleine Störungen ist, wie z. B. für Softwarefehler, Hardware- und Energieausfälle."[2]

Ein detailliertes Verständnis ist nicht nur für kritische Anwendungsdomänen wünschenswert, sondern für alle Domänen, in denen die Modelle eingesetzt werden. Die Studie bestätigt zudem, dass die größten Befürchtungen in Deutschland (auch gegenüber ausländischen Firmen) bezüglich eines Mangels an Nachvollziehbarkeit bei Ergebnissen einer KI bestehen [Del20]. Vor allem Negativbeispiele aus den Medien erhöhen die Forderung nach stärkerer Nachvollziehbarkeit solcher Modelle. Beim Autonomen Fahren verursachen die Systeme, die in der Regel auf tiefen neuronalen Netzen basieren, immer wieder Unfälle. Ein umfassendes Verständnis darüber, weshalb sich die Systeme in einer bestimmten Situation entsprechend verhalten haben, würde zur gezielten Fehlerbehebung und damit zu mehr Vertrauen in die Systeme führen. Gerade in Situationen, in denen ein autonomes Fahrzeug u. U. zwischen Leben und Tod entscheidet, sollen Systeme keine Blackbox sein.

Doch nicht nur Anwender fordern mehr Nachvollziehbarkeit. Gesetzlich wird versucht, Betroffenen eine rechtliche Grundlage zu bieten. Die automatisierte Entscheidungsfindung kann u. U. bestimmen, ob eine Person eine bestimmte Versicherung, ein Vorstellungsgespräch oder einen Kredit erhält. Die erneuerte EU-Datenschutz-Grundverordnung (DSGVO)[3] beschreibt in Artikel

[1] Zur besseren Lesbarkeit wird in der vorliegenden Arbeit auf die gleichzeitige Verwendung männlicher und weiblicher Sprachformen verzichtet. Es wird das generische Maskulinum verwendet, wobei beide Geschlechter gleichermaßen gemeint sind.

[2] [Hol20], S. 36.

[3] https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32016R0679, letzter Abruf am 30.12.2020.

13-15, dass Betroffene *aussagekräftige Informationen über die involvierte Logik* erhalten sollen [Hoe18]. Genauer wird für KI-Systeme vorgeschlagen, den Betroffenen ausschlaggebende Kriterien der Entscheidungsfindung offenzulegen [Hoe18].

Der Forschungsbereich zur Erzeugung nachvollziehbarer Modelle bzw. Modellergebnisse, der in den letzten Jahren sowohl in der Industrie als auch in der Forschungsgemeinschaft stark an Interesse gewonnen hat, wird als *Erklärbare Künstliche Intelligenz* (engl. XAI) benannt. Gunning [Gun17] führte den Begriff erstmals im Rahmen der Vorstellung eines gleichnamigen Forschungsprojektes ein, das von der Defense Advanced Research Projects Agency (DARPA) geleitet wurde. Die DARPA[4] ist eine Behörde des US-Verteidigungsministeriums, die Forschungsprojekte für die Streitkräfte durchführt. Das bereits abgeschlossene Projekt diente dazu, maschinelle Lernverfahren zu entwickeln, die es den Anwendern ermöglichen sollten, KI-Systeme zu verstehen, um diesen angemessen vertrauen zu können [Gun19].

Die Begriffe *Erklärbarkeit* und *Interpretierbarkeit* werden im Bereich der Erklärbaren KI oft synonym verwendet. Aufgrund der Vielzahl der Verfahren, die nahezu täglich publiziert werden, ist es zunehmend komplizierter geworden, die Verfahren entsprechend einem einheitlichen Modell einzuordnen. Eine weitere Herausforderung stellt die Tatsache dar, dass wenige Studien darüber existieren, inwieweit das Verständnis der Modelle oder bestimmter Modellergebnisse durch den Einsatz von Erklärungen verbessert wird.

Im nachfolgenden Abschnitt werden zunächst die Forschungsfragen der Arbeit beschrieben. Darauf aufbauend werden die Beiträge der Arbeit entsprechend der Forschungsfragen zugeordnet. Anschließend wird ein Überblick über den Aufbau der Arbeit gegeben.

[4] https://www.darpa.mil/, letzter Abruf am 10.01.2021.

1.1 Forschungsfragen der Arbeit

Im Folgenden werden die Forschungsfragen erläutert, die im Rahmen der vorliegenden Arbeit untersucht werden. Diese beschreiben offene Arbeitspunkte im Teilbereich des erklärbaren überwachten maschinellen Lernens.

Welche Arten von Erklärungen können im Bereich des überwachten maschinellen Lernens unterschieden werden?

Ein essenzieller Punkt im Themenbereich des erklärbaren maschinellen überwachten Lernens ist die Klassifizierung und Einordnung unterschiedlicher Arten von Erklärungen. So ist es von elementarer Bedeutung, eine einheitliche Grundlage dafür zu schaffen, welche Art von Erklärung entsprechend einem vorliegenden Vorhersagemodell generiert werden kann. Dazu ist die Untersuchung dessen, welche Arten von Erklärungen es gibt und wie diese – basierend auf einem vorliegenden Vorhersagemodell wie einer Blackbox (bspw. ein tiefes neuronales Netz) – erzeugt werden können, grundlegend. Durch diese Klassifizierung soll der Themenbereich des erklärbaren maschinellen Lernens strukturiert werden. Weiter ist es möglich, bei fehlender Nachvollziehbarkeit eines vorliegenden Vorhersagemodells eine geeignete Art der Erklärung genau zu spezifizieren, um mit dieser das Modell näher zu untersuchen. Der Beitrag dieser Arbeit zur Beantwortung der Fragestellung wird in Abschnitt 1.2.1 beschrieben, in dem ein theoretisches Vorgehensmodell eingeführt wird, das unterschiedliche Arten von Erklärungen einordnet und klassifiziert.

Wie können sowohl lokale als auch globale Erklärungen aus Blackbox-Modellen extrahiert werden, ohne dabei die Genauigkeit der Modelle zu beeinflussen?

Eine der Herausforderungen des erklärbaren maschinellen Lernens ist es, die globale Nachvollziehbarkeit basierend auf einer Blackbox zu generieren, ohne dabei die Genauigkeit des Modells negativ zu beeinflussen. Surrogat- oder auch sog. Stellvertreter-Modelle ahmen die Entscheidungsgrenzen einer Blackbox nach und erzeugen dadurch nachvollziehbare Entscheidungen für

den Anwender. Die Herausforderungen bestehen dabei darin, die Nachvollziehbarkeit zu generieren, eine hohe Modelltreue gegenüber der Blackbox zu bewahren und die Genauigkeit der Blackbox nicht zu beeinträchtigen.

Eine weitere Herausforderung des erklärbaren maschinellen Lernens stellt die Generierung von lokalen Erklärungen dar. Im Gegensatz zu globalen Erklärungen beziehen sich die Erklärungen auf eine einzelne oder wahlweise auch mehrere einzelne Instanzen. Eine Instanz bildet dabei einen Datenpunkt mit beliebig vielen Dimensionen ab.

Neben dem lokalen Aspekt liegt der Fokus der Fragestellung vor allem darauf, wie eine Erklärung für eine gegebene Instanz generiert werden kann, die möglichst nahe einer Entscheidungsgrenze liegt. Die Anforderung eine Erklärung zu finden, die möglichst nahe zur ursprünglichen Instanz liegt, resultiert daraus, da diese für eine Vielzahl von Anwendungen (z. B. bei Kreditanfragen) essenziell sein kann. Sofern ein Kunde, dessen Kreditanfrage abgelehnt wurde, einen neuen Kreditantrag stellen möchte, ist es hilfreich, die Gründe für die Ablehnung zu kennen und zu erfahren, unter welchen Bedingungen eine Bewilligung möglich ist. Eine Erklärung, die an einer Entscheidungsgrenze liegt, kann dem Anwender darüber Aufschluss geben, welche Merkmale mindestens geändert werden müssen um eine Kreditbewilligung zu erhalten. Die Beiträge der Arbeit zur Beantwortung dieser Fragestellung werden in Abschnitt 1.2.2 beschrieben.

Inwiefern wird durch die Bereitstellung von lokalen und globalen Erklärungen die Nachvollziehbarkeit der Modelle bzw. einzelner Modellergebnisse für Anwendergruppen verbessert?

Der Ruf nach Erklärungen im Bereich des erklärbaren maschinellen Lernens sowohl in der Industrie als auch in der Wissenschaft ist groß. Es gibt jedoch wenig Transparenz darüber, wie und ob lokale und globale Erklärungen die Nachvollziehbarkeit für betroffene Anwender verbessern können. Im Hinblick auf ein Vorhersagemodell, das eine Entscheidung erzeugt, soll eine Erklärung die Ursachen für die Entscheidung derart darstellen, dass diese von einem Anwender nachvollzogen bzw. verstanden werden kann. Dies

steht in direktem Zusammenhang mit dem Maß an Vertrauen, das diesem Modell entgegengebracht wird. Sachverhalten, die nicht verstanden werden, wird kein Vertrauen entgegengebracht. Daher ist es notwendig, im Rahmen von Benutzerstudien zu untersuchen, inwieweit erzeugte lokale und globale Erklärungen eine verbesserte Nachvollziehbarkeit generieren. Hierfür ist auch das Hintergrundwissen der Anwender zu berücksichtigen.

Einen weiteren Punkt bildet die Untersuchung im Bereich expliziter Merkmale. Explizite Merkmale können beispielsweise durch die Einbindung von Ontologien erstellt werden. Dabei entstehen aus den ursprünglichen Merkmalen neue Merkmale, die diese zusammenfassen. Der Einfluss expliziter Merkmale auf die Erklärung von Vorhersagemodellen wurde bislang nicht ausreichend untersucht. Vorhersagemodelle existieren in der Regel, um Anwender in unterschiedlichen Domänen zu unterstützen. Daher gilt es zu untersuchen, welche Art von Merkmalen in einer gegebenen Erklärung bevorzugt und besser verstanden wird. Der Beitrag der Arbeit zu dieser Fragestellung wird in Abschnitt 1.2.3 beschrieben.

1.2 Eigene Beiträge

Die Forschungsfragen, die in Abschnitt 1.1 erläutert wurden, werden durch die nachfolgend beschriebenen Beiträge der Arbeit aufgegriffen. Diese Beiträge waren bereits Inhalt von Veröffentlichungen, die in Abb. 1.1 aufgeführt sind. Die Forschungsbereiche der Arbeit lassen sich in zwei Kategorien unterteilen. Die erste Kategorie beinhaltet theoretische Erkenntnisse (s. Kapitel 3) und Verfahren (s. Kapitel 4 und 5). Die Verfahren wurden auf Grundlage der theoretischen Erkenntnisse abgeleitet. In der zweiten Kategorie wird die praktische Anwendung der theoretischen Erkenntnisse und Verfahren im Rahmen von Nutzerstudien untersucht.

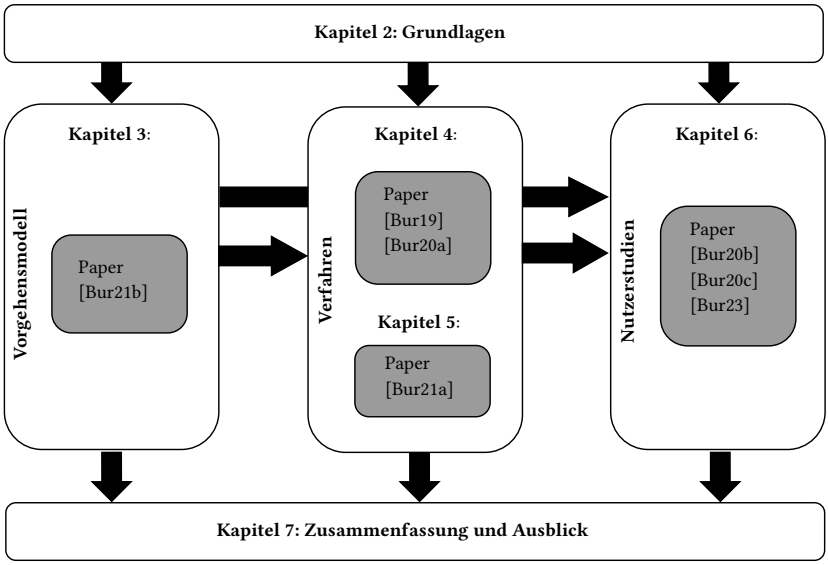


Abbildung 1.1: Überblick über die Beiträge [Quelle: Eigene Darstellung].

In Abb. 1.1 wird der Zusammenhang der Arbeit – basierend auf den unterschiedlichen Forschungsbereichen und Veröffentlichungen – dargestellt.

1.2.1 Vorgehensmodell zur Erzeugung unterschiedlicher Arten von Erklärungen

Dieser Teil der Arbeit basiert auf der Veröffentlichung von Burkart et al. [Bur21b]. Das Vorgehensmodell unterscheidet, welche Arten von Erklärungen im Bereich überwachter ML (vorwiegend Klassifikation) extrahiert werden können. Insgesamt werden fünf verschiedene Arten von Erklärungen unterschieden. Diese können dabei zunächst grob in Modell- und Instanz-Erklärungen unterteilt werden. Modell-Erklärungen erzeugen für das gesamte Modell auf globaler Ebene eine Erklärung, sodass der Anwender

jede Entscheidung des Modells nachvollziehen kann. Diese Art der Erklärungen wird auch globale Erklärung genannt. Instanz-Erklärungen beziehen sich darauf, ein bestimmtes Ergebnis oder ggfs. mehrere Ergebnisse eines Modells lokal nachvollziehen zu können. Diese werden auch als lokale Erklärungen bezeichnet. Bei Modell-Erklärungen wird zwischen drei Arten unterschieden: Direkte Modell-Erklärungen, Whitebox-Modell-Erklärungen und Modell-Erklärungen basierend auf einem globalen Surrogat-Modell. Instanz-Erklärungen unterscheiden sich nach direkten Instanzerklärungen sowie nach Instanzerklärungen basierend auf einem lokalen Surrogat-Modell.

1.2.2 Verfahren zur Erzeugung von Erklärungen

Dieser Teil der Arbeit basiert auf den Veröffentlichungen von Burkart et al. [Bur19] und Burkart et al. [Bur20a]. Insgesamt sind zwei Verfahren entworfen worden, die zum einen die globale und zum anderen die lokale Erklärbarkeit des Modells bzw. einzelner Instanzen unterstützen sollen.

Das erste Verfahren erzeugt ein Surrogat in Form einer Regelliste, dass als globale Erklärung für ein Modell dienen sollen. In einer Vorstufe basiert das Verfahren auf den Arbeiten von Wu et al. [Wu18] und Faller [Fal19]. Darauf aufbauend wurden Optimierungen durchgeführt. Zunächst wird als Hauptmodell ein neuronales Netz trainiert und daraus ein Surrogat (Regelliste) mithilfe der Regularisierung erzeugt. Die Länge dieser Regelliste wird als Strafterm auf die Verlustfunktion des neuronalen Netzes addiert. Das Ziel ist eine hohe Genauigkeit zu erreichen und dennoch die Nachvollziehbarkeit des Modells zu erzeugen.

Das zweite Verfahren erzeugt ein lokales Surrogat-Modell, das nahe zu einer Entscheidungsgrenze der ursprünglichen Dateninstanz liegt. Das Ziel ist es, die lokale Entscheidungsgrenze für die Dateninstanz zu finden und eine Darstellung der Entscheidungsgrenze mit einem nachvollziehbaren Modell, dem lokalen Surrogat, zu erstellen. Teil dieses Verfahrens ist die Erzeugung von kontrafaktischen Dateninstanzen, sodass dabei auch kontrafaktische Erklärungen erzeugt werden können.

1.2.3 Untersuchungen zur Nachvollziehbarkeit von Erklärungen mit Anwendern

Dieser Teil der Arbeit basiert auf den Veröffentlichungen von Burkart et al. [Bur20c], Burkart et al. [Bur20b] und Burkart et al. [Bur23]. In diesem werden Nutzerstudien beschrieben, die durchgeführt wurden, um die Nachvollziehbarkeit von Erklärungen mit Anwendern zu untersuchen. Ein wesentlicher Punkt ist die Untersuchung der Frage, inwieweit die Nachvollziehbarkeit für Anwender durch die Bereitstellung einer Erklärung verbessert wird. Dazu wurden insgesamt drei Nutzerstudien mit unterschiedlichen Anwendergruppen durchgeführt.

Die erste Nutzerstudie untersuchte basierend auf einer Fragestellung im maritimen Bereich [Bur20b] mehrere Arten von Erklärungen mit einem Domänenexperten. Zu diesem Zweck wurde eine Benutzerstudie mit einem maritimen Offizier durchgeführt, der vier unterschiedliche Arten von Erklärungen im Rahmen einer Klassifikationsaufgabe bewertete.

Die zweite Nutzerstudie befasste sich mit zwei Arten von Erklärungen, die sowohl lokale als auch globale Erklärungen enthielten. Die Versuchsteilnehmer waren Endbenutzer ohne weiteres domänenspezifisches Fachwissen. Hier wurde analysiert, welche Art der Erklärung mehr Nachvollziehbarkeit bietet [Bur20c].

Die dritte Nutzerstudie widmete sich der Fragestellung, inwieweit Erklärungen mit expliziten Merkmalen, d. h. kombinierten Merkmalen, die beispielsweise durch den Einsatz von Ontologien erreicht werden können, für Anwender besser nachvollziehbar sind [Bur23].

Die Resultate der Benutzerstudien zeigten, dass unabhängig von der Anwendergruppe Erklärungen die Nachvollziehbarkeit von Blackbox-Modellen steigern können. Für Domänenexperten eignen sich globale Erklärungen, da diese in der Regel den gesamten Entscheidungsprozess und nicht nur Teilaspekte des Entscheidungsprozesses verstehen wollen. Für Endbenutzer ohne speziell gefordertes Fachwissen eignen sich lokale Erklärungen, da für diese die vorliegende Entscheidung, jedoch nicht der gesamte Entscheidungsprozess der

Blackbox von Interesse war. Die Darstellung der Merkmalsrelevanz wurde von beiden Anwendergruppen gleichermaßen nachvollzogen und präferiert. Bei den Benutzerstudien war die Nachvollziehbarkeit der einzelnen Merkmale ausschlaggebend. Die Untersuchungen zeigten, dass explizite Merkmale besser verständlich sind. Die detaillierten Merkmale wurden in schwerwiegenden Situationen (wie sie z. B. im medizinischen Bereich auftreten) bevorzugt. Außerdem stellte sich heraus, dass die detaillierten Merkmale bevorzugt wurden, wenn die expliziten leicht zu erschließen sind.

1.3 Gliederung der Arbeit

Die Arbeit umfasst fünf inhaltliche Kapitel. Das Kapitel 2 gibt zunächst eine Übersicht über den Stand von Forschung und Technik. Dieser beschreibt einführend die Themenbereiche des maschinellen Lernens mit dem Aspekt der Erklärbarkeit. Weiterhin werden elementare Konzepte wie beispielsweise die Regularisierung vorgestellt, die für ein entworfenes Verfahren verwendet wird. Anschließend werden Basisverfahren zur Erzeugung von Erklärungen beschrieben. In Kapitel 3 wird sodann das theoretische Vorgehensmodell zur Klassifizierung unterschiedlicher Arten von Erklärungen vorgestellt, das das Rahmenwerk der vorliegenden Arbeit bildet. Insgesamt werden fünf unterschiedliche Arten von Erklärungen beschrieben. Das Kapitel 4 widmet sich anschließend der Entwicklung eines Verfahrens zur Extraktion eines globalen Surrogat-Modells basierend auf einem Blackbox-Modell. Hierfür wird zunächst die Problemdefinition erarbeitet. Ausgehend davon werden relevante Vorarbeiten im Bereich der Regularisierung erläutert, bevor die Methodik zur Problemlösung basierend auf der Regularisierung vorgestellt wird. Abschließend wird eine Evaluation des Verfahrens vorgenommen. Das Kapitel 5 beschreibt daraufhin die Entwicklung eines Verfahrens zur Extraktion eines lokalen Surrogat-Modells basierend auf einem Blackbox-Modell. Hierfür liegt der Schwerpunkt zuerst wieder auf der Problemdefinition. Ausgehend davon wird die Methodik zur Problemlösung beschrieben, die ein phasenweises Vorgehen beinhaltet. Im Anschluss folgt eine Evaluation des vorgeführten Verfahrens. In Kapitel 6 werden der Aufbau, die Durchführung und die Ergebnisse

der Benutzerstudien zur Bewertung unterschiedlicher Arten von Erklärungen erläutert. Insgesamt werden drei Benutzerstudien vorgestellt. Den Schluss der Arbeit bildet das Kapitel 7, das die wesentlichen Ergebnisse der Arbeit zusammenfasst und einen Ausblick auf mögliche weiterführende Forschungen gibt.

2 Stand von Forschung und Technik

Dieses Kapitel gibt einen Überblick über den Stand von Forschung und Technik zur Erklärbarkeit im Kontext maschineller Lernverfahren. Zunächst wird dazu eine Übersicht über die verschiedenen Lernparadigmen gegeben. Im weiteren Verlauf des Kapitels wird der Fokus auf das in der Arbeit relevante Lernparadigma, dem überwachten maschinellen Lernen, basierend auf tabellarischen Daten gelegt. Anschließend wird das Gebiet des erklärbaren überwachten maschinellen Lernens eingeführt. In diesem Zuge werden grundlegende Definitionen gegeben und die Taxonomie und essenzielle Eigenschaften von Erklärungen betrachtet. Darauf folgend wird die Regularisierung, die für das Kapitel 4 von Bedeutung ist, beschrieben. Abschließend wird ein Überblick über Basisverfahren gegeben, die auf Konzepten des überwachten Lernens basieren. Auf Grundlage dieser Verfahren beruhen die meisten Verfahren zur Erzeugung von Erklärungen. Bekannte Verfahren zur Erzeugung von Erklärungen werden direkt den Abschnitten der Basisverfahren zugeordnet. Die Beiträge dieses Kapitels sind in Anlehnung an Burkart et al. [Bur21b] aufgebaut.

2.1 Maschinelle Lernverfahren

Maschinelle Lernverfahren, insbesondere das Teilgebiet des tiefgehenden Lernens (engl. Deep Learning), erzielten in den letzten Jahren immer größere Erfolge. Die Verfahren lernen aus der Erfahrung, die in Form von historischen Daten vorliegt. Je nach Datengrundlage und dem zu lösenden Problem lassen

sich maschinelle Lernverfahren in unterschiedliche Lernparadigmen einteilen. Dazu wird in den folgenden Kapiteln ein kurzer Überblick über die unterschiedlichen Arten gegeben. Das Lernparadigma des überwachten maschinellen Lernens – vorallem der Klassifikation –, das im Fokus der vorliegenden Arbeit steht, wird ausführlich in Kapitel 2.1.2 erläutert. Abschließend werden die Metriken zur Bewertung von Klassifikatoren beschrieben.

2.1.1 Überblick über Lernparadigmen

In der nachstehenden Abb. 2.1 werden die unterschiedlichen Lernparadigmen dargestellt. Darin wird zwischen wissensbasierten und datengetriebenen Ansätzen, die auch *Human-in-the-Loop*-Ansätze enthalten, unterschieden. Aus der Abbildung geht hervor, dass die Notwendigkeit der Erklärbarkeit mit dem Anstieg der Komplexität der unterschiedlichen Lernparadigmen wächst. Systeme, die zu Beginn vom Anwender selbst entwickelt wurden (wie bspw.. durch regel- oder wissensbasierte Ansätze), galten als nachvollziehbar. Bei wissensbasierten Ansätzen wird die Wissensbasis vom Anwender aufgebaut. Datengetriebene Ansätze entkoppeln das System und den Anwender. Der Anwender wird als eine passive Komponente gesehen, der die Daten für die Trainingsalgorithmen vorbereitet, z. B. bei der Kennzeichnung der Daten mit Klassenbezeichnern für die Klassifikation. Human-in-the-Loop-Ansätze beziehen den Anwender als aktive Komponente mit ein, da der Lernalgorithmus aktives Feedback zur Lösung der geforderten Aufgaben verlangt.

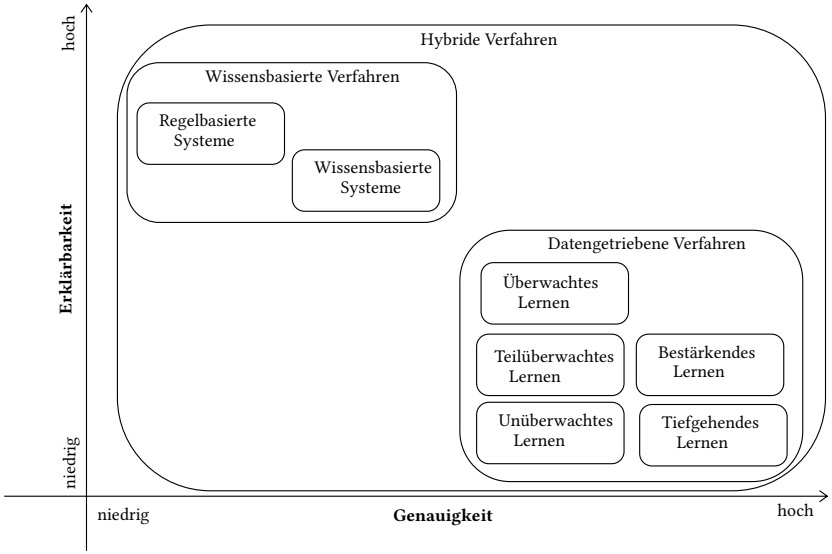


Abbildung 2.1: Qualitative Einordnung von Lernansätzen im Kontext der Erklärbarkeit [Quelle: Eigene Darstellung].

2.1.1.1 Wissensbasierte Ansätze

Wissensbasierte Ansätze beschreiben Verfahren, bei denen das domänenspezifische Wissen innerhalb einer Wissensbasis gespeichert wird, um bestimmte Probleme, wie z. B. zur Diagnostik einer bestimmten Krankheit in der Medizin, lösen zu können. Ein wichtiger Aspekt dabei ist die Trennung zwischen der Wissensbasis und der Wissensverarbeitung. Die Wissensbasis enthält spezifisches Wissen über ein Gebiet, während es sich bei der Wissensverarbeitung um eine anwendungsunabhängige Komponente handelt. Die frühen Formen fundierten auf regelbasierten Systemen, bei denen die Wissensrepräsentation auf vordefinierten, von Menschen generierten Regeln beruhte. Regeln werden als formalisierte Sätze in der Form *Wenn A Dann B* ausgedrückt. Solche Systeme waren für Menschen leicht zu verstehen, weil sie das Wissen dahinter als einfache Regeln definierten.

In der nächsten Stufe wurden Expertensysteme implementiert. Die Idee war es, das Expertenwissen in einer Wissensbasis zusammenzufassen, um die Maschine auf einem bestimmten Gebiet intelligent werden zu lassen. Über die Zeit hinweg stellte sich heraus, dass es eine anspruchsvolle und zeitintensive Aufgabe ist, das Wissen der Experten zu sammeln und entsprechend aufzubereiten. Für Experten eines bestimmten Gebietes kann es komplex sein, ihr Wissen – wie bspw. bei einer speziellen Behandlung einer Krankheit – zu formalisieren. Der Experte kann u. U. basierend auf seiner langjährigen Erfahrung die Entscheidung für eine entsprechende Behandlung treffen, die vom klassischen Vorgehen abweicht.

2.1.1.2 Datengetriebene Ansätze

Für den Bereich der datengetriebenen Ansätze werden nachfolgend die unterschiedlichen Lernparadigmen des maschinellen Lernens unterschieden.

Unüberwachtes Lernen (Unsupervised Learning): Das unüberwachte Lernen versucht, eigenständig Muster innerhalb der Daten zu finden, Anomalien zu erkennen oder die Dimensionalität der Daten zu reduzieren. Eine Trainingsinstanz besteht aus einem Merkmalsvektor, der im Gegensatz zum überwachten Lernen keine Zielvariable besitzt. Unüberwachte Lernalgorithmen wie bspw. das *Clustering* gruppieren die Trainingsdaten auf Grundlage der Ähnlichkeit der Trainingsinstanzen zu bestimmten Gruppen (Cluster). Neue Dateninstanzen werden basierend auf den Ähnlichkeitsmaßen einer Gruppe zugeordnet. Ein klassisches Beispiel dieser Anwendung findet sich beim Online-Kauf in der Sektion *Kunden, die diesen Artikel gekauft haben, kauften auch folgende Artikel* [Kra19]. Die bekanntesten Verfahren neben dem bereits erwähnten Clustering sind die Assoziationsanalyse und die Dimensionsreduktion. Ziel der Assoziationsanalyse ist es, Regeln innerhalb der Daten aufzufinden, die eine starke Korrelation aufweisen. Bei der Dimensionsreduktion geht es darum, die Vielzahl an vorhandenen Merkmalen des Datensatzes auf wenige essenzielle Merkmale zu beschränken.

Teilüberwachtes Lernen (Semisupervised Learning): Das Lernparadima des teilüberwachten Lernens verwendet sowohl Daten mit als auch ohne Zielvariable, um Vorhersagen zu treffen. Aufgrund des hohen Aufwandes für das Kennzeichnen von Daten mit der Zielvariable verwendet das teilüberwachte Lernen nur eine kleine Menge von Daten mit Zielvariable zusammen mit einer großen Menge an Daten ohne Zielvariable. Das teilüberwachte Lernen wird vor allem in der Objekt- und Bilderkennung angewendet. Ein bekanntes Verfahren aus diesem Bereich ist die *Label-Propagation-Methode*. Dabei werden mit dem Clustering die Daten in Gruppen aufgeteilt. In jedem Cluster kann den Daten ohne Zielvariable die entsprechende Zielvariable zugeordnet werden, die die anderen Daten innerhalb des Clusters aufweisen. Das aktive Lernen ist ein weiteres Lernparadigma, das dem teilüberwachten Lernen zugeordnet wird. Ein Anwender wird hierbei interaktiv aufgefordert, für bestimmte Daten die Zielvariable einzutragen.

Bestärkendes Lernen (Reinforcement Learning): Dieses Lernparadigma beruht auf dem Grundsatz von Belohnung und Bestrafung. In diesem Rahmen wird die Interaktion zwischen einem Agenten und dem Lernalgorithmus verwendet, um das Lernverhalten eines Agenten durch die Bestärkung zu untersuchen. Das Ziel ist die Optimierung einer sogenannten *Belohnungsfunktion*, indem Feedback von einem Agenten gegeben wird. Für jede Aktion erhält der Agent entweder eine positive oder eine negative Belohnung (Bestrafung). Durch die ständige Wiederholung von Situationen merkt sich das System die Aktionen mit den positiven Belohnungen. Mithilfe der Implementierung eines Randomisierungsfaktors kann der Raum möglicher Aktionen erkundet und so verhindert werden, dass der Algorithmus in einem lokalen Optima stecken bleibt [Lor20].

Interaktives Lernen (Interactive Learning): Das interaktive Lernen ist eines der neuesten Lernparadigmen. Holzinger bezeichnet diesen Ansatz als „Algorithmen, die mit – teils menschlichen – Agenten interagieren und durch diese Interaktion ihr Lernverhalten optimieren können“[5]. Es kombiniert das

[5] [Hol16], S. 64.

bestärkende Lernen, das aktive Lernen und das Online-Lernen und wird als eine Untermenge der *Human-in-the-Loop-Lernalgorithmen* angesehen [Hol16]. Interaktiv werden die Modelle durch die Eingabe der Anwender aufgebaut. Die Anwender können die Ergebnisse überprüfen, Korrekturen am Lernalgorithmus vornehmen und dem Lernalgorithmus dabei Feedback geben. Dies kann je nach Lernalgorithmus in nur einer oder mehreren Iterationen erfolgen.

2.1.1.3 Hybride Verfahren

Datengetriebene Verfahren wurden in den letzten Jahren intensiv zur Lösung vieler Probleme in den Bereichen der Bildverarbeitung, der Verarbeitung natürlicher Sprache und der Spracherkennung eingesetzt. Ein großer Nachteil dieser Verfahren ist jedoch die immense Datenenge, die diese für das Training benötigen um eine hohe Genauigkeit zu erreichen. Bengio [Ben19] merkt ein weiteres Problem der datengetriebenen Verfahren an, nämlich dass diese schwach darin sind, wenn diese über die Trainingsverteilung hinaus verallgemeinern sollen. Hybride Verfahren kombinieren wissensbasierte und datengetriebene Verfahren und können bei beiden erwähnten Problemen unterstützen. So können durch die datengetriebenen Verfahren Muster extrahiert werden, die anschließend durch wissensbasierte Verfahren angepasst bzw. angereichert werden können. Durch eine Anreicherung der Daten mit Expertenwissen werden meist auch weniger Trainingsdaten benötigt. Die Kombination der beiden Verfahren steigert auch die Erklärbarkeit der Verfahren da das Expertenwissen mit verarbeitet wird.

2.1.2 Überwachtes Lernen (Supervised Learning):

Beim überwachten Lernen werden annotierte Daten verwendet, um Zielvariablen für neue Instanzen vorhersagen zu können. Die Trainingsdaten bestehen aus den Merkmalen $\mathcal{X}$ und den Zielvariablen $\mathcal{Y}$. Die neuen Daten bestehen aus Merkmalen für die Zielvariablen vorhergesagt werden soll. Das Ziel beim überwachten maschinellen Lernen ist es, ein Modell $h(x) = y$ zu erlernen, das einen Merkmalsvektor $x \in \mathcal{X}$ auf ein Ziel $y \in \mathcal{Y}$ abbildet. Zu diesem

Zweck wird eine Menge von Trainingsdaten $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ zum Trainieren des Modells verwendet.

Algorithmen des überwachten Lernens werden in zwei verschiedene Lernaufgaben unterteilt: die Klassifikation und die Regression. Im Falle der Klassifikation besteht das Ziel darin, einen diskreten Wert y vorherzusagen, der in der Regel als *Label* bezeichnet wird. Bei einer binären Klassifikation ist das Label bspw. $y \in \{0, 1\}$. Die Aufgabe der Regression besteht hingegen darin, mit $y \in \mathbb{R}$ eine kontinuierliche Zielvariable vorherzusagen.

Beim überwachten maschinellen Lernen wird im Allgemeinen versucht, die Verlustfunktion in Gleichung (2.1) zu optimieren, wobei $\mathcal{D}$ der Trainingsdatensatz und h^* das resultierende Modell ist. Das Maß S gibt den Fehler der zu optimierenden Zielfunktion an. Die Verlustfunktion bestimmt die Differenz zwischen der Vorhersage des Modells und einem vorgegebenen Klassenbezeichner.

Problem 1 (Überwachtes maschinelles Lernen)**.**

$$h^* = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} S\big(h(x_i), y_i\big)\,, \tag{2.1}$$

mit $(x_i, y_i) \in \mathcal{D}$.

2.1.2.1 Klassifikation

Nachfolgend wird die weitere Beschreibung der Problemdefinition ausschließlich auf die Klassifikation gerichtet. Generell wird zwischen zwei Arten von Modellen unterschieden, die das Resultat der Klassifikation sein können. Das Modell kann einerseits eine Blackbox $b : \mathcal{X} \to \mathcal{Y}$, $b(x) = y$ mit $b \in \mathcal{B}$ sein, indem $\mathcal{B} \subset \mathcal{H}$ der Hypothesenraum der Blackbox ist bspw. kann b ein neuronales Netz mit einer versteckten Verbindungsschicht sein. Andererseits kann es sich um ein Whitebox-Modell $w : \mathcal{X} \to \mathcal{Y}$, $w(x) = y$ mit $w \in \mathcal{W}$ handeln, indem $\mathcal{W} \subset \mathcal{H}$ der Hypothesenraum des interpretierbaren

Whitebox-Modells ist bspw. kann w ein Entscheidungsbaum der Tiefe drei sein.

Um die Klassifikationsleistung eines trainierten Modells zu bewerten, wird das Maß S verwendet. Ein wichtiges Beispiel für ein Leistungsmaß in der binären Klassifikation ist das Maß der *Genauigkeit* mit $S = \frac{1}{n}\sum_i |h(x_i) - y_i|$, das das Verhältnis aller richtig klassifizierten Instanzen von $\mathcal{D}$ darstellt. Sofern ein Leistungsmaß gegeben ist, kann das Problem des überwachten maschinellen Lernens als Optimierungsproblem formuliert werden. Die Aufgabe der Optimierung ist es, die Verlustfunktion zu optimieren bzw. den Fehler zu minimieren.

Essenziell ist die Unterscheidung zwischen dem *Lernalgorithmus* und dem resultierenden Modell: Der Lernalgorithmus versucht, das Optimierungsproblem aus Gleichung (2.1) direkt oder implizit zu lösen. Das Ergebnis des Optimierungsproblems und damit auch die Ausgabe des Lernalgorithmus ist das eigentliche Modell, das sodann auf eine neue Dateninstanz x angewendet werden kann, um eine Vorhersage $y = h^*(x)$ zu generieren.

Im Falle eines parametrischen Modells mit dem Parametervektor $h(x;\theta)$ wie bspw. einem neuronalen Netz wird Gleichung (2.1) äquivalent formuliert als

$$\theta^* = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} S\big(h(x_i;\theta), y_i\big) \,. \tag{2.2}$$

Oft lässt sich das Optimierungsproblem aus Gleichung (2.1) oder Gleichung (2.2) nicht exakt analytisch lösen. Eine der wenigen Ausnahmen, bei denen dies möglich ist, ist die lineare Regression. Daher ist es üblich, dass numerisch eine suboptimale Lösung gefunden wird, wie am Beispiel von tiefen neuronalen Netzen, bei denen der Parametervektor θ mittels eines Gradientenabstiegsverfahrens bestimmt wird.

Martens et al. [Mar11] unterscheiden Algorithmen der Klassifikation auf Grundlage des Ausgabetyps. Dabei werden die am häufigsten verwendeten Typen eingeführt, die sich wie folgt unterteilen lassen: lineare Modelle,

nichtlineare Modelle, regelbasierte Modelle, baumbasierte Modelle, Nächste-Nachbarn-Modelle und Bayes'sche Netzwerke. Beispiele für nichtlineare Modelle sind Künstliche Neuronale Netze (KNN) oder Stützvektormaschinen (SVM). Die Abbildung der Funktion, die verwendet wird, um ein Ergebnis basierend auf dem Merkmalsvektor zu erhalten, ist meist von nichtlinearer Natur. Die Modelle können für die Ein-Klassen- oder die Multiklassen-Klassifizierung verwendet werden. Die Ein-Klassen-Klassifizierung lernt, Objekte einer bestimmten Klasse unter den gesamten Objekten zu identifizieren. Die Multiklassen-Klassifizierung lernt, zwischen mehreren Klassen zu unterscheiden. Im Folgenden werden bekannte Verfahren aus dem Bereich der Klassifikation beschrieben. Zunächst werden jedoch nur diejenigen erläutert, die als Blackbox gelten. In Abschnitt 2.5 werden Verfahren der Klassifikation beschrieben, die aufgrund ihrer Eigenschaften als nachvollziehbare Modelle gelten.

Support Vector Machine (SVM): SVMs werden für binäre Klassifikationsprobleme verwendet, die entweder linearer oder nicht linearer Natur sein können. Das Ziel ist es, die Eingabemerkmale in einem anderen Merkmalsraum abzubilden, sodass die Entscheidungsgrenze den Raum zwischen den beiden möglichen Klassen maximiert. Um Daten nichtlinearer Natur zu klassifizieren, wird der sogenannte *Kernel-Trick* angewendet. Dabei werden die Eingabemerkmale in einem hochdimensionalen nichtlinearen Merkmalsraum abgebildet. Anstatt diese Abbildung zu berechnen, verwendet das Verfahren einen im Allgemeinen nichtlinearen Kernel. Die Entscheidungsgrenze ist eine trennende Hyperebene. Diese wird durch die Stützvektoren jeder Klasse bestimmt. Die Linearkombination definiert diese trennende Hyperebene. Ein Stützvektor stellt dabei einen Datenpunkt aus dem Datensatz dar, der die trennende Hyperebene stützt und damit der Hyperebene am nächsten liegt [Fla12]. Diese Stützvektoren werden für die Berechnung der Entscheidungen verwendet, wodurch die Komplexität dieser Berechnungen reduziert wird. SVMs gelten als effizient in hochdimensionalen Merkmalsräumen.

Bayes'sche Netze: Die Netze stellen die Wahrscheinlichkeiten von Ereignissen und deren Abhängigkeiten zueinander dar. Diese verwenden

das Bayes'sche Theorem sowie vordefinierte Abhängigkeiten zwischen Merkmalen, um die wahrscheinlichste Klasse für eine neue Dateninstanz vorherzusagen [Koc13]. Für die Klassifizierung werden alle Merkmale verwendet. Bayes'sche Netze erweitern den Naiven-Bayes-Klassifikator, der keine Abhängigkeiten zwischen den Eingabemerkmalen annimmt. Der Naive Bayes-Klassifikator kann mittels Boosting (s. Abschnitt Ensemble-Verfahren) zu einem Ensemble-Verfahren erweitert werden.

Ensemble-Verfahren: Zur Kombination mehrerer Entscheidungsbäume z. B. stehen verschiedene Konzepte zur Verfügung, sogenannte Ensemble-Verfahren, sodass ein Random Forest (RF) entsteht. Ensemble-Verfahren kombinieren mehrere Klassifikatoren der gleichen oder auch unterschiedlicher Art, um ein genaueres Vorhersageergebnis zu erreichen [Fla12]. Die Verfahren zeichnen sich durch ihre erhöhte Komplexität hinsichtlich der Kombination der einzelnen Vorhersagemodelle und des gelernten finalen Modells aus. Die Vorhersagemodelle werden in der Regel aus zufälligen Teilmengen der Daten oder unter Verwendung zufälliger Teilmengen der Eingabemerkmale konstruiert. Es gibt verschiedene Möglichkeiten diese zu kombinieren. Beispielsweise indem die individuellen Vorhersagen gemittelt werden. Wahlweise können die Vorhersagemodelle auch sequenziell aufgebaut werden, sodass die Gesamtverzerrung reduziert wird und zu einer besseren finalen Vorhersage führen kann. Eine Kombination mehrerer Vorhersagemodelle durch die Berechnung des Mittelwerts der individuellen Vorhersagen kann die gesamte Vorhersagegenauigkeit erhöhen. Eine weitere Möglichkeit stellt die Mehrheitsentscheidung dar. Die bekanntesten Methoden zur Kombination von Ensemble-Verfahren sind *Bagging*, *Boosting* oder *Stacking*. Das *Bagging* verwendet *Bootstrap-* und *Subraum*-Stichproben, um eine größere Vielfalt in den Lernprozess der einzelnen Vorhersagemodelle einzuführen. Bootstrap-Stichproben sind zufällige Stichproben des Datensatzes. Subraum-Stichproben berücksichtigen diejenigen Merkmale, die vom Vorhersagemodell verwendet werden. Eine zufällige Teilmenge der verfügbaren Merkmale wird verwendet, um jedes einzelne Vorhersagemodell zu trainieren. Beim *Boosting* werden mehrere Modelle der gleichen Art sequenziell trainiert, sodass Fehler, die von einem Vorhersagemodell in einem Schritt des Lernprozesses erzeugt wurden, im nächsten Schritt kompensiert

werden können. Dies kann durch die Einführung von Stichprobengewichten erreicht werden. Wird eine Stichprobe richtig klassifiziert, wird ihr Gewicht reduziert. Wird eine Stichprobe jedoch falsch klassifiziert, wird ihr Gewicht erhöht. Zusätzlich wird jedem einzelnen Vorhersagemodell ein Gewicht zugewiesen, das die Sicherheit der Vorhersage angibt. Dieses Gewicht wird zur Berechnung der endgültigen Vorhersage verwendet, die als gewichtete Summe der einzelnen Vorhersagen berechnet wird. Sofern ein Klassifikator einen erhöhten Fehler aufweist, werden die falsch klassifizierten Eingabedaten die Wahl des nächsten Klassifikators stärker prägen. Das *Stacking* ähnelt dem *Boosting*. Hierbei wird ein Meta-Modell gelernt, das die verschiedenen Vorhersagemodelle, die einzeln gelernt wurden, kombiniert.

Künstlich Neuronale Netze (KNN): Das Gehirn enthält eine Vielzahl von Neuronen, die über Synapsen miteinander verbunden sind. Diese Verbindungen ermöglichen es, Signale effizient zu übertragen. Sobald ein *Neuron* ein oder mehrere Signale empfängt, sendet dieses selbst ein Signal aus. Durch diesen Vorgang entsteht ein Netzwerk von Informationssignalen. Künstlich Neuronale Netze (nachfolgende nur noch Neuronale Netze genannt) adaptieren dieses Konzept. Künstliche Neuronen sind über Kanten verbunden, die Signale in Form von numerischen Werten transportieren können. Jedes dieser Neuronen kann mehrere solcher Signale empfangen und berechnet über eine sogenannte Aktivierungsfunktion, wann genau ein Signal gesendet werden soll. In der Literatur gibt es verschiedene Arten von Aktivierungsfunktionen, so z. B. Rectified Linear Unit (ReLu), Sigmoid oder Softmax. Bei der Aktivierungsfunktion ReLu werden alle negativen Werte auf Null gesetzt, wobei die Neuronen deaktiviert werden. Dieses Vorgehen birgt u. a. große Geschwindigkeitsvorteile bei der Berechnung.

Künstliche Neuronen sind in verschiedenen Schichten aufgebaut. Die Gewichte an den Kanten beeinflussen das Signal, das auf dieser Verbindung transportiert wird. Durch die Änderung der Gewichte wird ein Netzwerk derart ausgeprägt, dass es eine bestimmte Aufgabe erfüllen kann. Ein neuronales Netz mit vielen versteckten Schichten wird als tiefes neuronales Netz bezeichnet. Es gibt verschiedene Arten von neuronalen Netzen wie das

Multilayer-Perzeptron (MLP), vorwärtsgerichtete, faltende oder rekurrente neuronale Netze. Das MLP ist eine Ergänzung der vorwärtsgerichteten Netze und kann aus beliebig vielen Schichten $\mathcal{N}$ bestehen. Die Arten der Schichten unterteilen sich in die Eingangsschicht, die versteckten Schichten und die Ausgangsschicht. Die Eingangsschicht erhält das zu verarbeitende Eingangssignal. Die Klassifizierung wird durch die Ausgabeschicht durchgeführt. Die Neuronen sind über Gewichte θ_n mit $n \in 1,...,N$ verbunden, die in einer Gewichtsmatrix $\Theta = \{\theta_n\}_{n=1}^{N}$ abgelegt sind. Eine beliebige Anzahl von versteckten Schichten, die zwischen der Eingabe- und der Ausgabeschicht angeordnet sind, führen die Berechnungen durch. Ähnlich wie bei den vorwärtsgerichteten Netzen fließen die Daten vorwärtsgerichtet von der Eingabe- zur Ausgabeschicht. Die Neuronen werden mit der Fehlerrückführung (engl. Backpropagation) trainiert. Mehrschichtige Perzeptronen erlauben die Approximation beliebiger kontinuierlicher Funktionen und können Probleme lösen, die nicht linear trennbar sind.

2.1.2.2 Regression

Die lineare Regression wird verwendet, um Vorhersagen für kontinuierliche Werte wie z. B. Gehalt oder Alter zu erzeugen. Hierzu wird eine lineare Beziehung zwischen einer oder mehreren unabhängigen Variablen und einer abhängigen Variablen angesetzt:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_z x_z + s. \tag{2.3}$$

Das Ergebnis einer Instanz ist die gewichtete Summe der gesamten Merkmale. Die Gesamtheit der β stellen die gelernten Merkmalsgewichte dar, wobei β_0 als Schnittpunkt mit der y-Achse (engl. Intercept) bezeichnet und mit keinem Merkmal multipliziert wird. s stellt den Fehler dar, d. h. die Differenz zwischen der Vorhersage und dem tatsächlichen Ergebnis. Ein gutes Regressionsmodell generiert genaue Vorhersagen im Bereich der trainierten Ausgabewerte. Sobald die neue Dateninstanz außerhalb der konvexen Hülle der Trainingsdaten

liegt, muss das Regressionsmodell extrapolieren. Dies stellt ein unerwünschtes Verhalten dar. Grundsätzlich kann zwischen der simplen linearen Regression und der multiplen linearen Regression unterschieden werden. Die Verwendung einer einzelnen unabhängigen Variable zur Vorhersage des Wertes einer numerischen abhängigen Variable wird als einfache lineare Regression bezeichnet. Wird mehr als eine unabhängige Variable verwendet, um den Wert einer abhängigen Variable vorherzusagen, handelt es sich um eine multiple lineare Regression. Regressionsverfahren können auch nichtlinearer Natur sein.

2.1.3 Bewertung von Klassifikatoren

Zur Bewertung eines binären Klassifikators können unterschiedliche Metriken angewendet werden: Accuracy, Precision, Recall und der F1-Score. Zu Beginn des Modelltrainings werden die Daten z. B. mit der k-fachen Kreuzvalidierung in einen Test- und einen Trainingsdatensatz aufgeteilt. Die Metriken werden anhand des Testdatensatzes ermittelt. Die Genauigkeit ist definiert als der Anteil der korrekt klassifizierten Testinstanzen. Die *Precision* gibt die Wahrscheinlichkeit an, mit der das Vorhersagemodell eine Klassifizierung korrekt durchführt. Der *Recall* gibt an, wie gut ein Vorhersagemodell eine Instanz richtig klassifiziert. Die Precision und der Recall können unter Verwendung des harmonischen Mittels kombiniert werden, um den F_1-Score zu bilden. Die Konfusionsmatrix, die der Abb. 2.2 entnommen werden kann, visualisiert die Beziehungen zwischen den vorhergesagten und den tatsächlichen Ergebnissen. Aus der Konfusionsmatrix lassen sich die Formeln für die Metriken ableiten. Die Konfusionsmatrix wird zudem als Interpretationswerkzeug für wissenschaftliche Ergebnisse verwendet [Ros20]. Auch diese wird im Rahmen der Erklärbarkeit eingesetzt. Die Untersuchung der Konfusionsmatrix ermöglichte z. B. die Identifizierung von bioakustischen Ähnlichkeiten zwischen verschiedenen Arten von Fröschen [Col18]. Anhand der Formeln (s. Gleichung 2.4-2.7) können diese Metriken berechnet werden.

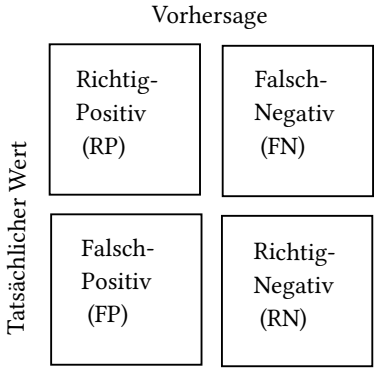


Abbildung 2.2: Konfusionsmatrix [Quelle: Eigene Darstellung].

$$\text{Accuracy} = \frac{\text{RP} + \text{RN}}{\text{RP} + \text{FN} + \text{FP} + \text{RN}}. \tag{2.4}$$

$$\text{Precision} = \frac{\text{RN}}{\text{RN} + \text{FP}}. \tag{2.5}$$

$$\text{Recall} = \frac{\text{RN}}{\text{RN} + \text{FN}}. \tag{2.6}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.7}$$

Ein weiteres Mittel zur Bewertung von Klassifikatoren ist die Receiver Operating Characteristic (ROC)-Kurve oder die Area Under The Curve (AUC). In

Abb. 2.3 ist eine Darstellung abgebildet, die bei der Unterscheidung unterstützen soll. Die ROC-Kurve ist eine Darstellung in der die Richtig-Postiv-Rate gegen die Falsch-Positiv-Rate dargestellt wird. Ein sehr guter Klassifikator würde alle positiven Objekte korrekt klassifizieren und eine Falsch-Positiv-Rate von 0 % vorweisen. Die Fläche unter der ROC-Kurve wird AUC genannt. Der AUC-Wert gibt an wie gut der Klassifikator zwischen Klassen unterscheiden kann. Je höher der AUC Wert ist, desto eher wird ein positiver Wert auch als solcher klassifiziert [Nar18] und desto weniger ein negativer fälschlicherweise positiv.

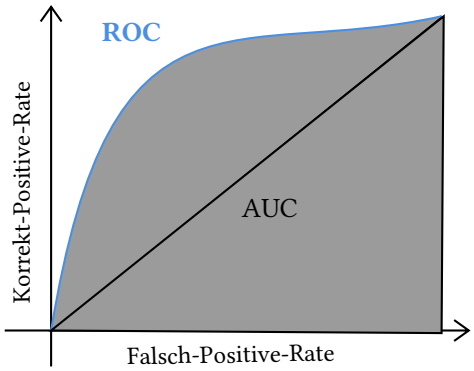


Abbildung 2.3: Darstellung der ROC-Kurve mit AUC [Quelle: In Anlehnung an [Nar18]].

2.2 Erklärbarkeit maschineller Lernverfahren

Maschinelle Lernverfahren werden auf Daten trainiert, um spezielle Fragestellungen zu lösen. In vielen Fällen handelt es sich dabei um historische Daten, die zum Beispiel menschliche Entscheidungen enthalten. Basierend auf diesen historischen Daten wird gelernt, eine Entscheidung zu treffen. Dazu trainiert

der jeweilige Algorithmus mithilfe der vorliegenden Daten ein Modell, das zur Vorhersage verwendet wird.

Historische Daten enthalten nicht selten Fehler, die zu falschen Ergebnissen führen können. Das durch die maschinellen Lernverfahren trainierte Vorhersagemodell wird solche Fehler aufgreifen und fehlerhafte Entscheidungen treffen, sofern diese nicht aufgedeckt werden. Beispielsweise kann eine Bank historische Daten zu Kreditentscheidungen ihrer Kunden, die diese über Jahre hinweg abgefragt und gesammelt hat, dazu nutzen, um ein Modell zu trainieren. Dieses kann sodann zur automatisierten Bewertung von Kreditvergaben verwendet werden. Fehler lassen sich z. B. in zufällige und systematische Fehler unterscheiden. Sofern in den Daten ein zufälliger Fehler, wie zum Beispiel eine fehlerhafte nicht vergebene Kreditanfrage, enthalten ist, lernt das Modell diese mit und trifft die fehlerhafte Entscheidung u. U. weiterhin. Systematische Fehler sind Verzerrungen, die sich auf eine systematische Abweichung beziehen.

In der Vergangenheit war beim maschinellen Lernen fast ausschließlich eine hohe Accuracy (s. Gleichung (2.4)) der Modelle von großer Bedeutung. Ein unbeachteter Aspekt war die Erklärbarkeit der Modelle. Dabei gab es frühe Arbeiten, wie die Merkmalsrelevanz von Random Forests [Bre01a], die als Meilenstein in diesem Bereich gilt [Mol20]. Über die Zeit hinweg zeichnete sich vermehrt ab, dass die Erklärbarkeit der Modelle entscheidend für den Erfolg der Anwendungen in unterschiedlichen Bereichen ist. In der Literatur finden sich viele Beispiele, in denen das Modell zwar eine hohe Genauigkeit aufweist, jedoch auf Grundlage falscher Zusammenhänge die Entscheidung trifft [Rib16b].

Ein Beispiel für eine Verzerrung aus dem medizinischen Bereich wurde von Caruana et al. [Car15] beschrieben. In der Arbeit wurde ein System zur Vorhersage der Sterblichkeit von Patienten mit einer Lungenentzündung trainiert. Dafür wurden unterschiedliche Modelle basierend auf unterschiedlichen Verfahren trainiert. Genauer gesagt wurden die Ergebnisse von neuronalen Netzen und regelbasierte Modelle untersucht und verglichen. Die regelbasierten Modelle wurden in diesem Experiment vor allem aufgrund ihrer Nachvollziehbarkeit verwendet. Das regelbasierte Modell lernte eine Regel,

die besagte, dass Patienten mit einer Lungenentzündung, die an Asthma litten, ein geringeres Risiko haben, an einer Lungenentzündung zu sterben, als diejenigen ohne Asthma. Die Regel spiegelte ein wahres Muster in den Trainingsdaten wider, war jedoch fehlerhaft. Patienten, die an Asthma litten und mit einer Lungenentzündung ins Krankenhaus eingeliefert wurden, wurden direkt intensivmedizinisch betreut und erhielten zu Beginn eine aggressive Behandlung. Bedingt durch dieses Vorgehen hatten die Patienten, die im Vorhinein eine aggressivere Behandlung erhielten, ein geringeres Risiko, an einer Lungenentzündung zu sterben. Durch die historischen Daten trainierten die Modelle die Regel, dass Asthma das Risiko senkt, obwohl Asthmatiker ein deutlich höheres Risiko hatten, an einer Lungenentzündung zu sterben.

Das Forschungsfeld der *Erklärbaren Künstlichen Intelligenz* versucht u.a. solche Arten von Verzerrungen frühzeitig sichtbar zu machen. Generell wird versucht sowohl bestimmte Entscheidungen als auch gesamte Modelle, die basierend auf maschinellen Lernverfahren trainiert wurden, für den Anwender verständlich und nachvollziehbar werden.

In den folgenden Abschnitten werden zunächst Gründe und Domänen beleuchtet, in denen die Nachvollziehbarkeit, die durch das Konzept der Erklärbarkeit umgesetzt wird, von Bedeutung ist. Anschließend werden elementare Definitionen und Begriffe eingeführt. Darauf aufbauend wird die Taxonomie der Erklärbarkeit erläutert, die für die weitere Einordnung der Arbeit grundlegend ist.

2.2.1 Gründe für die Forderung nach Erklärbarkeit

Systeme, die basierend auf trainierten Modellen automatisierte Entscheidungen treffen (wie z. B. in der Medizin), werden meist nicht direkt von den Anwendern akzeptiert, da diese mehr Einblicke in die Entscheidungsfindung haben wollen [Štr10]. Vor allem wenn diese Entscheidung die Anwender persönlich betrifft – wie z. B. bei einer abgelehnten Kreditanfrage. Vertrauen in die Systeme aufzubauen, ist einer der wichtigsten Gründe bei der Forderung nach Erklärbarkeit. Das Vertrauen in das Modell ist für die Einführung innerhalb einer bestimmten Anwendungsdomäne deshalb zwingend erforderlich.

Das Verständnis und die Kenntnis der Stärken und Schwächen des Vorhersagemodells sind eine Grundvoraussetzung für das Vertrauen der Anwender und damit für den Einsatz des Vorhersagemodells. Die *Generalisierung* des Modells ist ein weiterer Grund. Das Vorhersagemodell sollte Anwendern zukünftiges Verhalten vermitteln, damit das Modell vertrauensvoll mit neuen Daten verwendet werden kann. Nur wenn sichergestellt ist, dass das Modell gut verallgemeinert, bzw. zumindest bekannt ist, in welchem Kontext es gut verallgemeinert, kann das Vorhersagemodell für eine automatisierte Entscheidungsfindung eingesetzt werden [Lip18]. Die Kenntnis der Gründe für eine bestimmte automatisierte Entscheidung ist ein gesellschaftliches Bedürfnis, um sicherzustellen, dass eine faire Entscheidung getroffen wurde. Auch könnte dies als offizielles Recht für EU-Bürger entsprechend der DSGVO ausgelegt werden [Goo17]. Entscheidungsträger könnten dadurch rechtlich dazu aufgefordert werden, die Entscheidungen für betroffene Personen nachvollziehbar zu gestalten, um ethische Richtlinien einzuhalten.

Erklärbarkeit wird nicht explizit in jeder Anwendungsdomäne verlangt. Die Anwendungsdomänen, in denen Erklärbarkeit nicht benötigt wird, sind entweder gut erforscht, sodass die Anwender den vorhandenen Modellen vertrauen oder keine schwerwiegenden Konsequenzen zu erwarten sind, falls das System Fehler macht – bspw. bei Empfehlungssystemen für den Online-Einkauf [Dos17]. Nach Lipton [Lip18] wird Erklärbarkeit immer dann gefordert, wenn das Ziel, für das das Modell konstruiert wurde, von der tatsächlichen Anwendung, für die das Modell eingesetzt wird, abweicht. Der Bedarf an Erklärbarkeit entsteht also aufgrund einer Diskrepanz zwischen dem, was ein Modell erklären kann, und dem, was ein Anwender wissen will. Nach [Mar09] ist Erklärbarkeit immer dann wichtig, wenn ein Modell validiert werden muss, bevor es implementiert und eingesetzt werden kann. Bereiche, die Erklärbarkeit erfordern, zeichnen sich dadurch aus, dass diese kritische Entscheidungen treffen, die bspw. Menschenleben betreffen [Štr10].

2.2.2 Erklärbarkeit innerhalb kritischer Infrastrukturen

Der Einsatz von KI in Bereichen der kritischen Infrastrukturen wie z. B. dem Bevölkerungsschutz birgt großes Potenzial [Koh20]. „Kritische Infrastrukturen (KRITIS) sind Organisationen oder Einrichtungen mit wichtiger Bedeutung für das staatliche Gemeinwesen, bei deren Ausfall oder Beeinträchtigung nachhaltig wirkende Versorgungsengpässe, erhebliche Störungen der öffentlichen Sicherheit oder andere dramatische Folgen eintreten würden“[6]. Hierzu gehören folgende Domänen: Energie, Gesundheit, Staat und Verwaltung, Ernährung, Transport und Verkehr, Finanz- und Versicherungswesen, Informationstechnik und Telekommunikation, Medien und Kultur und Wasser. Die Voraussetzung für den Einsatz innerhalb kritischer Infrastrukturen ist die Vertrauenswürdigkeit der Systeme. Dabei bildet das Konzept der Erklärbarkeit einen essenziellen Baustein zur Generierung von Vertrauen in die Systeme. Biran et al. [Bir17] sind der Ansicht, dass Menschen einer Vorhersage durch Modelle der KI nur dann vertrauen werden, wenn das System die getroffene Entscheidung rechtfertigen kann. Sunyaev [Sun20] stellt heraus: „Selbst wenn ein KI-Modell erfolgreich trainiert wurde, stellt die Erklärbarkeit dieses Modells eine weitere Herausforderung dar: Wenn ein KI-Modell zum Beispiel eine bestimmte Therapie empfiehlt, aber weder der Patient noch der Arzt oder ein Informatikexperte nachvollziehen können, wie diese Empfehlung zustande gekommen ist, fehlen Arzt und Patient gleichermaßen das Vertrauen in das KI-System und sie könnten daher von der Anwendung der KI-basierten Empfehlung absehen.“[7] Im Folgenden werden Beispiele für kritische Bereiche gegeben, in denen das Konzept der Erklärbarkeit von Bedeutung ist.

Legal Tech: Der Anwendungsbereich, in dem KI in der Justiz angewendet wird, nennt sich *Legal Tech*. Dieser umfasst alle Belange, die sich mit Technik und Recht beschäftigen [Kes20]. Auch wenn aktuell nach Artikel 92 in Verbindung mit Artikel 97 des Grundgesetzes eine Rechtsprechung nur durch eine natürliche Person erfolgen darf [Bia21], kann in Zukunft zumindest eine

[6] https://www.bbk.bund.de/DE/AufgabenundAusstattung/KritischeInfrastrukturen/kritischeinfrastrukturen_node.html, letzter Abruf am 09.01.2021.

[7] [Sun20], S. 111- S. 112.

automatisierte Unterstützung durch KI bei der Urteilsfindung erfolgen. IBM Watson bietet mit *Justiz Memoria*[8] bereits einen ersten Demonstrator für den juristischen Bereich des Mietrechts. Hierzu werden Textinhalte automatisch analysiert und wichtige Informationen aus ähnlichen Verfahren hinzugezogen.

Autonomes Fahren: Ist ein autonom fahrendes Auto in einen Unfall verwickelt, ist es für den Entwickler, für die beteiligten Personen und die Justiz wichtig, die Ursachen für den Unfall detailliert zu verstehen. Dabei gilt es festzustellen, ob das System fehlerhaft gehandelt hat oder nicht. Auch das Vertrauen in die Fahrzeuge ist essenziell für ihren Erfolg. Gerade durch Unfälle in diesem Bereich wie bspw. bei Uber [Lev18] oder Tesla [Pol18] entstehen Vorbehalte gegenüber der Verwendung, die u. a. durch den Einsatz von Erklärbarkeitsverfahren gemildert werden können.

Medizin: Microsofts *InnerEye* bietet eine grafische Benutzeroberfläche für Algorithmen, die Radiologen bei der Diagnose von Tumoren und der Planung präziser chirurgischer Eingriffe unterstützen. Das Unternehmen *DeepMind Health* will Modelle zur Diagnose gängiger Netzhautpathologien auf der Grundlage von optischen Tomografie-Scans entwickeln [Wat19]. Als Paradebeispiel von KI im Bereich der Medizin gilt IBMs *Watson* für die Onkologie. Watson soll dazu verwendet werden, um Therapieempfehlungen zu generieren. Allerdings wurden in der Vergangenheit nicht korrekte und nicht nachvollziehbare Therapieempfehlungen gegeben, sodass die Onkologen den Ergebnissen nicht vertrauten [Ros18]. Sofern die Onkologen verstanden hätten, wie Watson die Ergebnisse generiert, wären die Akzeptanz und das Vertrauen vermutlich größer gewesen.

Finanzwesen: Im Finanzwesen sind Prozesse oftmals sehr papierlastig. Einfache Aufgaben wie ein Datenabgleich im Rahmen einer Bonitätsprüfung könnten durch die Systeme effizient übernommen werden. Ebenso für die Berechnung eines Ausfallrisikos bei der Kreditvergabe oder bei neuen Kreditanfragen für die Einstufung der Kreditwürdigkeit anhand einer vorangegangenen

[8] https://www.ibm.com/de-de, letzter Abruf: 22.01.2021.

Analyse können die Verfahren vielversprechend eingesetzt werden. Auch hier ist jedoch die Nachvollziehbarkeit der Ergebnisse essenziell.

2.2.3 Grundbaustein der Erklärbarkeit – die Datenqualität

Die Qualität der Daten, mit denen die Modelle trainiert werden, ist maßgeblich in Bezug auf die Erklärbarkeit der maschinell gelernten Modelle. Erklärungen basieren auf den Vorhersagen eines Modells und sind somit von den Daten abhängig, die das Modell für seine Vorhersagen verwendet. Ist die Qualität der Daten bereits unzureichend, so wird die Vorhersage des Modells u. U. auch fehlerhaft sein. Modelle können aus unterschiedlichen Gründen unzureichende Vorhersagen basierend auf den falschen Trainingsdaten treffen:

- zu wenige Daten → stochastischer Fehler
- nicht repräsentative Daten →Verzerrung, systematischer Fehler
- fehlerhafte, widersprüchliche Daten → Inkonsistenz
- Störungen, Rauschen → stochastischer Fehler

Die korrespondierende Erklärung kann unter Umständen dennoch korrekt sein, beruht aber auf den falschen Daten. Sofern die Qualität der Daten bereits zu Beginn Fehler aufweist, wird sich dies im weiteren Verlauf der Datenverarbeitung immer weiter verschlechtern und verschleiern.

Aufschlussreich ist in diesem Zusammenhang ein Blick auf die Begriffsdefinition der Qualität. Eine zentrale Stelle für die Definition von Begriffen ist die Internationale Organisation für Normung[9]. Diese definiert Qualität wie folgt:

> „Vermögen einer Gesamtheit inhärenter (lat. innewohnend) Merkmale eines Produkts, eines Systems oder eines Prozesses zur Erfüllung von Forderungen von Kunden und anderen interessierten Parteien“[10].

[9] https://www.iso.org/home.html, letzter Abruf am 31.12.2020.
[10] [Kel08], S. 153.

Die Qualität wird demnach als Maß beschrieben, das widerspiegelt, wie viele der geforderten Anforderungen erfüllt werden. Kaggle[11], eine Online-Plattform für den Wissensaustausch und Wettbewerbe rund um das Thema Datenanalyse, führte eine Umfrage [Kag17] durch. Die Umfrage enthielt u.a. Fragen dazu, was die größte Barrieren eines Datenanalysten bei der täglichen Arbeit sind. Das Resultat der Umfrage war, dass die größte Barriere, der ein Datenanalyst bei der täglichen Arbeit gegenübersteht, die schlechte Qualität der vorliegenden Daten ist. Die größte Arbeit wird somit in die Bereinigung der Daten investiert. Sobald eine große Menge unvollständiger, veralteter, zu geringer und verrauschter Daten zum Training eines Modells verwendet wird, werden auch die resultierenden Modelle fehlerhaft sein. Daten können oft o.g. Fehler enthalten. In vielen Bereichen, wie bspw. im Gesundheits- oder Finanzwesen, in denen ein Modell Entscheidungen trifft, die Personen direkt betreffen, ist es zwingend erforderlich, dass diese Modelle nicht mit fehlerhaften Daten trainiert wurden. Einer der häufigsten Irrtümer ist es zu glauben, dass das Modell besser wird, je mehr Daten verfügbar sind. Peter Norvig, Forschungsdirektor des Unternehmens Google [Rus02], erklärt dazu, dass Google keine besseren Algorithmen als andere Unternehmen entwickelt hat, sondern nur mehr Daten besitzt. Der wesentliche Punkt dabei ist, dass es nicht ausreicht, über mehr Daten zu verfügen, sondern mehr konsistente Daten zu haben. Experimente, bei denen mehr konsistente Daten während des Trainings einbezogen wurden, zeigten, dass sich die Leistung des Modells nicht verbesserte [Ama17].

Gudivada et al. [Gud17] stellten weiter fest, dass die Verwendung qualitativ hochwertiger Daten viele positive Auswirkungen wie z. B. die Reduktion von Kosten haben kann. Wang et al. [Wan96] definieren Datenqualität als den Grad, in dem die verfügbaren Daten den Bedürfnissen des Anwenders entsprechen. Wie Helfert et al. [Hel16] aus dieser Definition abgeleitet haben, ist das Konzept der Datenqualität kontextabhängig und subjektiv. Die Dimensionen der Datenqualität lassen sich in sechs Kernbereiche aufteilen: *Vollständigkeit, Eindeutigkeit, Aktualität, Gültigkeit, Genauigkeit* und *Konsistenz* [Ask13]. Unter dem Begriff Vollständigkeit wird der Anteil der tatsächlich erhobenen

[11] https://www.kaggle.com/, letzter Abruf am 17.07.2021.

Daten im Vergleich zu dem Potenzial, das theoretisch hätte erhoben werden können, verstanden. Die Eindeutigkeit misst den Grad identischer Dateninstanzen. Die Aktualität definiert, dass die gesammelten Daten nicht zu alt sein dürfen. Die Gültigkeit prüft, ob die Syntax der Daten mit ihrer Definition übereinstimmt. Die Genauigkeit misst, ob die verfügbaren Daten die gegebene Aufgabe korrekt beschreiben. Die Konsistenz der Daten gibt darüber Aufschluss, ob die Daten alle bestimmte Anforderungen erfüllen.

2.2.4 Definition von Erklärbarkeit

Kodratoff [Kod94] definiert Erklärbarkeit grob als die Fähigkeit eines Modells, von den Anwendern verstanden zu werden. Die Definition von Erklärbarkeit im Kontext des maschinellen Lernens gilt als eine der größten Herausforderungen und ist bislang nicht einheitlich gelöst. Rudin [Rud18] führt aus, dass Interpretierbarkeit ein domänenspezifischer Begriff ist und es daher keine allgemeingültige Definition geben kann. Neben den unterschiedlichen Begriffen, wie Interpretierbarkeit oder Verständlichkeit, die neben der Erklärbarkeit meist synonym gebraucht werden, werden in der Literatur noch viele weitere verwendet.

Der Trend geht jedoch zur Unterscheidung der Begrifflichkeiten [Gui18a, Bur21b]. So beschreibt Interpretierbarkeit, dass das Modell von Natur aus nachvollziehbar ist, wie z. B. bei einem flachen Entscheidungsbaum. Erklärbarkeit beschreibt hingegen die Fähigkeit eines Blackbox-Modells, mithilfe externer Ressourcen (z. B. Visualisierungen) nachvollziehbar gestaltet zu werden [Bib21]. Murdoch et al. [Mur19] definieren die Bezeichnung *interpretierbares maschinelles Lernen* als „die Verwendung von Modellen des maschinellen Lernens für die Extraktion von relevantem Wissen über Domänenzusammenhänge, die in den Daten enthalten sind. Dabei wird Wissen als

relevant angesehen, wenn es ermöglicht, Einblicke innerhalb einer Domäne für eine bestimmte Zielgruppe zu erzeugen."[12]

Aktuell wird keine einheitliche Definition dafür gegeben, was genau Erklärbarkeit ist und wie diese standardmäßig gemessen werden kann. Dies ist aufgrund der subjektiven Eigenschaft, da diese u. a. von der Semantik und der Erfahrung der Anwender des Modells abhängen kann [Bib21], ein nicht triviales Problem. Molnar et al. [Mol20] erläutert, wie beurteilt werden soll, ob eine neue Methode Modelle besser erklärt, ohne eine zufriedenstellende Definition der Erklärbarkeit zu haben. Rechtliche Entscheidungsträger wie die EU-Kommission definieren die Erklärbarkeit sehr vage. Dort wird die Erklärbarkeit des algorithmischen Entscheidungsprozesses insofern beschrieben [Kom19], dass diese „so weit wie möglich gegeben werden sollte"[13]. Dabei wird nicht genau spezifiziert, wie diese gegeben werden soll bzw. wann genau diese erfüllt ist. Derzeit gibt es zwei übergeordnete Strategien zur Messung von Erklärbarkeit: Dazu gehört zum einen die objektive Bewertung über quantifizierbare Metriken und zum anderen über subjektive anwenderzentrierte Untersuchungen im Rahmen von Benutzerstudien. Beide Arten der Messung werden in Abschnitt 2.3.5 beschrieben.

In dieser Arbeit werden die Begriffe Erklärbarkeit und Interpretierbarkeit nicht synonym, sondern in Anlehnung an die Arbeiten von Burkart et al. [Bur21b] und Guidotti et al. [Gui18a] verwendet. Demnach ist Interpretierbarkeit auf natürliche Weise vom Modell gegeben. Der Begriff der Erklärbarkeit hingegen erweitert den Begriff der Interpretierbarkeit und wird verwendet, wenn Lösungen auf zusätzlichen Modellen, sogenannten Surrogaten, basieren oder direkt aus komplexen Modellen extrahiert werden. Somit umfasst der Begriff der Erklärbarkeit den Begriff der Interpretierbarkeit, wie auf der

[12] [Mur19], S. 22071– S. 22080, Eigene Übersetzung: „We define interpretable machine learning as the use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in data. Here, we view knowledge as being relevant if it provides insight for a particular audience into a chosen domain problem.".

[13] Eigene Übersetzung: [Bib21], S. 2 „...explainability of the algorithmic decision-making process, adapted to the persons involved, should be provided to the extent possible".

Abb. 2.4 dargestellt. Der Pfeil in der Abbildung von der Blackbox zur Whitebox symbolisiert, dass die Blackbox mit entsprechenden Verfahren zu einer Whitebox gestaltet werden kann. Die genaue Vorgehensweise dazu wird in Kapitel 3 beschrieben. Die Nachvollziehbarkeit bzw. die Verständlichkeit eines Modells wird durch die Verfahren der Erklärbarkeit erzeugt.

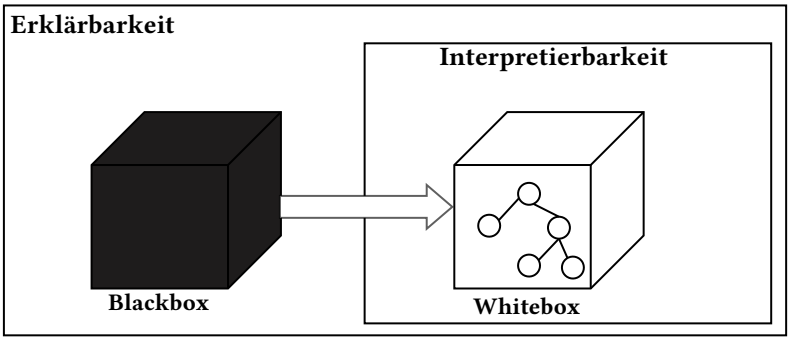


Abbildung 2.4: Interpretierbarkeit als Teilgebiet der Erklärbarkeit [Quelle: Eigene Darstellung].

2.2.5 Begriffe der Erklärbaren Künstlichen Intelligenz

Im Folgenden werden zunächst Begriffe, die im weiteren Verlauf der Arbeit von Bedeutung sein werden, erläutert. Zunächst werden die Bezeichnungen Vorhersagemodell, interpretierbares Vorhersagemodell (Whitebox-Modell), komplexes nicht nachvollziehbares Vorhersagemodell (Blackbox-Modell), das Surrogat- bzw. Stellvertreter-Modell sowie der Ausdruck Erklärung genauer eingeführt und terminologisch geklärt.

Vorhersagemodell oder Modell: Das Vorhersagemodell ist jenes Modell, das eine Vorhersage für eine neue Instanz trifft und mittels eines bestimmten Algorithmus trainiert wurde. Im Falle einer Klassifizierungsaufgabe ist das

Vorhersagemodell ein Klassifikator. Das Vorhersagemodell kann dabei entweder ein interpretierbares Vorhersagemodell oder ein komplexes Vorhersagemodell sein.

Interpretierbares Vorhersagemodell (Whitebox-Modell): Interpretierbare Vorhersagemodelle sind Modelle, die inhärent interpretierbar und somit direkt nachvollziehbar für den Anwender sind. Diese werden in ihrer Gesamtheit von der Anwendergruppe verstanden und auch als Whitebox-Modelle bezeichnet. Solche interpretierbaren Vorhersagemodelle können beispielsweise lineare Modelle, kleine Entscheidungsbäume, kurze Entscheidungslisten oder -tabellen sein. Die Vorhersagemodelle müssen einfach nachzuvollziehen sein und lassen sich vor allem mit niedrigdimensionalen Daten gut anwenden. Interpretierbarkeit gilt als die höchste Form des Verständnisses, die über ein Modell erlangt werden kann. Lipton [Lip18] nennt drei Kriterien, die interpretierbare Modelle erfüllen sollen.

- Simulierbarkeit: Die Anwendergruppe kann mithilfe der Eingangsdaten und der Modellparameter in angemessener Zeit jeden Berechnungsschritt einer Vorhersage verstehen.
- Algorithmische Transparenz: Der Lernalgorithmus, mit dem das Modell generiert wird, ist verständlich.
- Zerlegbarkeit: Die einzelnen Bestandteile wie die Eingabedaten, die Modellparameter oder die Berechnungen sind verständlich.

Ein Modell gilt als interpretierbar, wenn alle drei Kriterien erfüllt werden.

Komplexes Vorhersagemodell (Blackbox-Modell): Komplexe Modelle werden auch Blackbox-Vorhersagemodelle genannt. Blackbox-Modelle können aufgrund ihrer Komplexität nur schwer durch die Anwender nachvollzogen werden und zeichnen sich meist dadurch aus, dass lediglich die Eingabedaten und die Ausgabedaten (Ergebnisse) der Modelle bekannt und nachvollziehbar sind. Dabei ist nicht klar, wie das Vorhersagemodell die Ergebnisse abgeleitet hat. Bei der Generierung des Modells werden komplexe Eingabe-Ausgabe-Beziehungen – basierend auf den zugrundeliegenden Trainingsdaten – gelernt. Klassischerweise werden als Beispiel für komplexe

Vorhersagemodelle oft neuronale Netze genannt. Wenngleich die einzelnen Parameter des Netzes wie die Anzahl der Schichten, die Anzahl der Neuronen, die Aktivierungsfunktion und die gelernten Gewichte beim Lernen des Modells bekannt sind, ist die gelernte Funktion, die die Eingabe des Modells mit der Ausgabe in Beziehung setzt, zu kompliziert. Diese können demnach nicht einfach von den Anwendern nachvollzogen werden. Andere Beispiele für komplexe Vorhersagemodelle sind SVMs oder auch Ensemble-Methoden wie der Random Forest.

Surrogat- bzw. Stellvertreter-Modell: Ein Surrogat- oder Stellvertreter-Modell ist ein interpretierbares Modell, das ein komplexeres Modell wie die Blackbox versucht zu erklären. Um die Nachvollziehbarkeit zu erzeugen, werden beide Modelle – sowohl die Blackbox als auch das Surrogat-Modell – benötigt.

Erklärung: Eine Erklärung kann für ein spezielles Ergebnis oder ein gesamtes Vorhersagemodell gegeben werden und soll dem Anwender die Nachvollziehbarkeit ermöglichen. Dies kann der Fall sein, wenn das Vorhersagemodell sowohl eine Blackbox als auch eine Whitebox ist und zusätzliche Informationen über die Vorhersage gewünscht sind. Beispielsweise kann eine bestimmte Vorhersage einer Blackbox durch eine Erklärung ergänzt werden. Nach Guidotti et al. [Gui18a] kann eine Erklärung als eine Schnittstelle betrachtet werden, die die Kommunikation zwischen dem Anwender und dem Modell ermöglicht. Eine gute Erklärung zeichnet sich dadurch aus, dass diese für den Menschen nachvollziehbar ist, die Vorhersagen des Modells widerspiegelt und somit möglichst *modelltreu* ist [Lip18].

Instanz bzw. Dateninstanz:

Als Instanz wird ein Datenpunkt bezeichnet, der eine unterschiedliche Anzahl von Dimensionen entsprechend des Datensatzes enthalten kann. Die Dimension bezeichnet dabei die vorhandenen Merkmale einer Instanz. Beispielsweise kann eine Instanz ein bestimmter Kreditantrag einer Person mit den unterschiedlichen Merkmalen wie Beruf, Geschlecht oder Einkommen sein.

2.2.6 Taxonomie der Erklärbarkeit

In diesem Kapitel wird die Taxonomie der Erklärbarkeit näher betrachtet. Diese dient dazu, verschiedene Verfahren im Bereich der Erklärbarkeit einzuordnen. Sie bietet jedoch keine Strukturierung darüber, welche Art von Erklärungen extrahiert werden können.

Ante-hoc-Erklärbarkeit: Ante-hoc-Erklärbarkeit bedeutet, dass das Vorhersagemodell vorab darauf trainiert wird, nachvollziehbar zu sein. Somit wird der Aspekt der Erklärbarkeit in die Struktur des Vorhersagemodells eingearbeitet [Hol19] oder ist von Natur aus gegeben. Modelle, die im Hinblick auf die Erklärbarkeit ante hoc trainiert wurden, sind in der Regel interpretierbare Modelle. Es wird kein weiteres Modell benötigt, um die Nachvollziehbarkeit des Modells zu erzeugen.

Post-hoc-Erklärbarkeit: Post-hoc-Erklärbarkeit wird erst, nachdem ein Blackbox-Vorhersagemodell trainiert wurde, erzeugt. Der Aspekt der Erklärbarkeit wird somit im Nachgang betrachtet. Zunächst ist es das Ziel, ein möglichst genaues Vorhersagemodell zu trainieren, das beliebig komplex werden kann. Erst im Anschluss wird versucht, Erklärbarkeit zu erzeugen.

Lokale Erklärbarkeit: Die lokale Erklärbarkeit betrachtet die Entscheidung zu einer einzelnen Instanz [Phi18] und versucht, eine Erklärung dafür zu geben, warum eine spezifische Entscheidung getroffen wurde [Dos17]. Dies ist hilfreich, um eine bestimmte Entscheidung nachvollziehen und zugleich rechtfertigen zu können.

Globale Erklärbarkeit: Die globale Erklärbarkeit betrachtet das gesamte Modell und liefert ein Muster, das das Vorhersagemodell im Allgemeinen entdeckt hat [Phi18]. Das Verhalten eines Vorhersagemodells kann dem Anwender als Ganzes vermittelt werden, ohne dass explizit Vorhersagen einzelner Instanzen [Lak17] betrachtet werden. In der Literatur wird beschrieben, dass die globale Erklärbarkeit nur schwer erreicht werden kann, sobald das Vorhersagemodell komplex ist oder die Daten hochdimensional sind. Das Ziel der globalen Erklärbarkeit ist es, die allgemeinen Stärken, Schwächen und Grenzen des Vorhersagemodells aufzuzeigen und dem Anwender zu vermitteln,

wie sich dieses in einer unbekannten Situation verhält, um Vertrauen in das Vorhersagemodell aufbauen zu können [Wel17]. Ein globales Modellverständnis kann auch durch die Erklärung mehrerer repräsentativer einzelner Vorhersagen gegeben werden wie z. B. in den Arbeiten von Ribeiro et al. [Rib16b]. Modellerklärungen sind für ein fundiertes Verständnis und für die frühzeitige Erkennung von Verzerrungen essenziell [Dos17].

Modell-spezifisch (Spezifisch): Verfahren der Erklärbarkeit können entweder auf beliebige Modelle oder für spezielle Arten angewendet werden. Modell-spezifische Verfahren können nur auf eine bestimmte Art von Modell angewendet werden, da z. B. die interne Struktur des Vorhersagemodells von den Verfahren zur Erzeugung von Erklärungen verwendet wird. Diese Art von Verfahren gilt als wenig flexibel (im Gegensatz zu modell-agnostischen Verfahren), da diese auf eine spezielle Gruppe von Modellen beschränkt sind.

Modell-agnostisch (Agnostisch): Modell-agnostisch bedeutet, dass das Verfahren Erklärungen aus einer beliebigen Blackbox extrahieren kann [Rib16a]. Die Verfahren gelten als flexibel, da sie auf viele unterschiedliche Modelle angewendet werden können [Rib16a].

2.3 Erklärungen

Erklärungen geben die Antwort auf eine Warum-Frage [Mil19]. Miller [Mil19] geht auf die philosophische, psychologische und kognitive Sichtweise von Erklärungen ein. In diesem Zusammenhang erläutert er, dass das Gebiet der Erklärbaren KI ein Problem der Mensch-Maschine-Interaktion ist. Weiter zeigt er auf, dass Erklärungen anhand der folgenden Kriterien evaluiert werden können: Wahrscheinlichkeit, Einfachheit, Generalisierbarkeit und Übereinstimmung mit früheren Überzeugungen. Genauer wird beschrieben, dass der Aspekt der Wahrscheinlichkeit zwei Facetten hat: Zum einen die Wahrscheinlichkeit, zu der eine Erklärung wahr ist, und zum anderen die Angabe einer Wahrscheinlichkeit innerhalb einer Erklärung. Miller [Mil19] verweist auf die Theorie von Thagard [Tha89], die besagt, dass Menschen simple und generalisierte Erklärungen bevorzugen. Die menschliche Entscheidungsfindung kann

durch die Verwendung von *Generierung natürlicher Sprache* (engl. Natural Language Generation (NLG)) verbessert werden, insbesondere im Bereich unsicherer Daten [Gka16]. Im Folgenden wird auf die Desiderata von Erklärungen eingegangen, da es wichtig ist, diese vor der Erzeugung einer Erklärung festzulegen, um diese als Anforderung gezielt erfüllen zu können.

2.3.1 Desiderata von Erklärungen

Kass et al. [Kas88] stellen heraus, dass die Qualität einer Erklärung von drei unterschiedlichen Kriterien abhängig ist: der *Relevanz*, der *Überzeugungskraft* und der *Verständlichkeit* einer Erklärung. Eine Erklärung ist nur dann relevant, wenn sie den aktuellen Zielen und Bedürfnissen des Anwenders entspricht. Die Erklärung sollte möglichst viele Informationen enthalten, um diese Ziele zu erreichen. Zusätzlich sollte die Erklärung so kurz wie möglich sein, um zu vermeiden, dass Informationen gegeben werden, die nicht notwendig sind. Der Anwender ist von einer Erklärung überzeugt, wenn sie auf Fakten beruht, die dieser glaubt. Die Verständlichkeit einer Erklärung wird durch verschiedene Aspekte erreicht. So sollte die Erklärung einen bestimmten Darstellungstyp verwenden, die der Anwender leicht verstehen kann, prägnant sein und dem Anwender interessante Aspekte aufzeigen. Für ein System, das Erklärungen erzeugt, ist es wichtig, ein gewisses Maß an Flexibilität und Reaktionsfähigkeit zu besitzen. Versteht der Benutzer eine Erklärung nicht, sollte das System weiteres Wissen zur Verfügung stellen, um auf die speziellen Bedürfnisse des Benutzers zu reagieren [Swa91]. Fischer et al. [Fis90] bestätigen die These von Kass et al. [Kas88], dass eine Erklärung so kurz wie möglich sein sollte. Diese führen ein Erklärsystem ein, das ein Minimum an Erklärungen generiert. Erhält das System vom Anwender die Rückmeldung, dass die Erklärung nicht ausreichend war, fügt dieses der gegebenen Erklärung weitere Details hinzu [Fis90]. Dieser Ansatz ist in der Lage, die Bedürfnisse des Anwenders optimal zu befriedigen, ohne diesen dabei zu überfordern. Darüber hinaus versucht dieser Ansatz, komplexe Erklärungen zu vermeiden.

Hoffman et al. [Hof18] unternehmen ebenfalls den Versuch, festzulegen, was eine gute Erklärung ausmacht. Was für diese wichtig ist, ist subjektiv. Essenziell ist es zu wissen, welches Hintergrundwissen der Anwender bereits hat und welche Ziele er verfolgt. Dabei kann ein mentales Modell unterstützen, um nachvollziehen zu können, wie ein Anwender ein bestimmtes Ereignis, einen Prozess oder ein System versteht. Hoffman et al. [Hof18] stellen ein konzeptionelles Modell für die Erklärbare KI vor, welche ein mentales Modell enthält. Dieses kann auf der nachstehenden Abb. 2.5 entnommen werden.

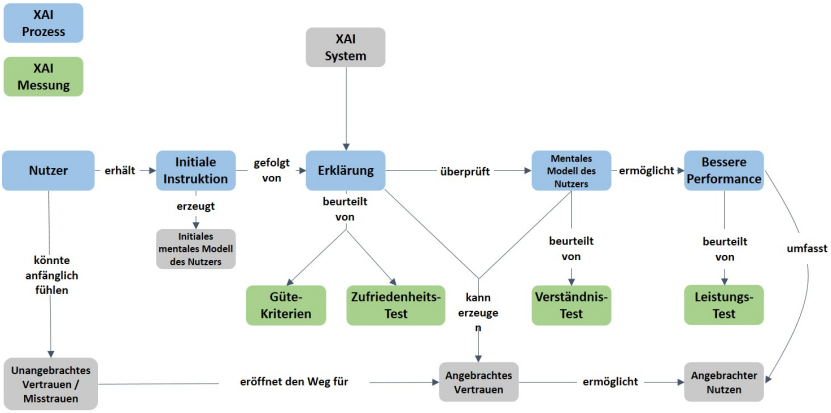


Abbildung 2.5: Konzeptionelles Modell des Erklärprozesses im XAI-Kontext [Quelle: In Anlehnung an [Hof18], S. 3].

Erklärungen, die gut sind und die Bedürfnisse des Anwenders erfüllen, tragen zu einer positiven Entwicklung des mentalen Modells bei. Dadurch vertraut der Nutzer dem System. Die erste Einweisung in die Nutzung eines Systems ermöglicht es dem Benutzer, ein erstes mentales Modell der Aufgabe und des Systems zu bilden. Die anschließende weiterführende Erfahrung, die auch die

vom System generierten Erklärungen beinhalten kann, würde es dem Teilnehmer ermöglichen, das mentale Modell zu verfeinern. Dies kann zu entsprechendem Vertrauen in das System führen. Auch die Berücksichtigung der Datenart ist für eine gute Erklärung von Relevanz. Das bedeutet, dass je nach Verfügbarkeit des Datentyps unterschiedliche Erkläransätze in Betracht gezogen werden müssen und sich die Art der Kommunikation ändern kann. Beispielsweise ist die Verwendung einer Heatmap zur Visualisierung des Gradienten eines Modells für Bilddaten nützlicher als für tabellarische oder textuelle Daten.

Lipton [Lip18] postuliert, dass sich Erklärungen auf anormale Gründe stützen sollten. Diese zeichnen sich dadurch aus, dass sie dem Anwender zunächst ungewöhnlich erscheinen. Zusätzlich betont Lipton [Lip18], dass ein nachvollziehbares Modell *menschlich-simulierbar* sein sollte. Hierunter ist zu verstehen, dass ein Anwender in der Lage sein sollte, die Ergebnisse eines Modells nachzuvollziehen. Der Anwender soll mit den Eingabedaten und den Parametern des Modells imstande sein, in angemessener Zeit jede Berechnung, die zur Erstellung einer Vorhersage erforderlich ist, auszuführen [Lip18].

2.3.2 Aufbau von Erklärungen

Für die Generierung einer Erklärung können die folgenden Leitfragen in Betracht gezogen werden[Bur21b]:

- Was soll erklärt werden (Inhaltsart)?
- Wie soll etwas erklärt werden (Kommunikationsweg)?
- Wem soll etwas erklärt werden (Anwendergruppe)?

2.3.2.1 Inhaltsart

Je nach Modell können unterschiedliche Arten von Erklärungen generiert werden. Um eine Entscheidung eines Modells oder ein gesamtes Modell zu erklären, muss eine bestimmte Art von Erklärung bzw. ein bestimmtes Stilelement gewählt werden. Darunter wird jedoch nicht z. B. die Visualisierung

der Erklärung verstanden. Diese ist Teil der Kommunikation. Als Inhalt einer Erklärung werden folgende Arten unterschieden:

Globale faktische Erklärung: Globale Erklärungen versuchen, das gesamte Modell nachvollziehbar zu machen. Die globale Erklärbarkeit bezieht sich auf die allgemeinen Strukturen [Phi18] und erzeugt ein Muster, das das Vorhersagemodell im Allgemeinen entdeckt hat. Das System kann das Verhalten eines Modells als Ganzes vermitteln, ohne sich auf Vorhersagen einzelner Instanzen zu beziehen [Lak17]. Diese Art der Erklärung wird ausführlich in Abschnitt 2.3.3 beschrieben.

Lokale faktische Erklärung: Diese Art der Erklärung bezieht sich auf eine einzelne Entscheidung und beschreibt den Grund (oder die Gründe) dafür, warum diese getroffen wurde [Dos17]. Lokale Erklärungen beziehen sich nur auf die umliegende Nachbarschaft einer Instanz, um die vorliegende Entscheidung zu erklären. Auf diese Art der Erklärung wird ausführlicher in Abschnitt 2.3.3 eingegangen.

Lokale kontrafaktische Erklärung: Dieser Typ der Erklärungen generiert die Möglichkeit, eine vorliegende Entscheidung zu verändern. Dabei werden Hinweise dafür gegeben, wie eine getroffene Entscheidung durch Änderung der gegebenen Merkmale verändert werden kann, um eine bevorzugte Entscheidung zu erreichen. Ein Beispiel, bei dem eine kontrafaktische Erklärung von Nutzen ist, wäre ein abgelehnter Kreditantrag. Der Kreditnehmer wäre in der Lage, bestimmte Anpassungen vorzunehmen, um so doch einen Kredit von der Bank zu erhalten. Diese Art der Erklärung wird ausführlicher in Abschnitt 2.3.3 beleuchtet. Globale kontrafaktische Erklärungen stellen aktuell ein offenes Forschungsgebiet dar.

Prototypische Erklärung: Rüping [Rüp06] beschreibt Prototypen als Beispiele, die einer großen Anzahl von Beispielen ähnlich sind. Diese können daher als Vertreter dieser Menge verwendet werden. Prototypische Erklärungen können dementsprechend durch das Geben von Erklärungen der jeweiligen Vertreter geliefert werden. Auch können nur ähnliche Prototypen zu der ursprünglichen Entscheidung gegeben werden. Beispiele sind prototypische

Instanzen, die der zu erklärenden Instanz ähnlich sind. Prototypische Erklärungen können sowohl lokal als auch global eingesetzt erstellt werden. Sofern nur der Prototyp für eine bestimmte Instanz gewählt wird, entspricht dies einer lokalen Erklärung. Werden alle Prototypen des Datensatzes als Erklärung dargestellt, zeigt dies eine globale Erklärung.

2.3.2.2 Die Art der Kommunikation

Die Kommunikationsart bestimmt, wie einem Anwender die Erklärung übermittelt wird. Dabei gibt es unterschiedliche Arten der Kommunikation, die entweder einzeln angewendet oder miteinander kombiniert werden können. Diese werden im weiteren Verlauf näher erläutert.

Textuelle Beschreibung: Die Erklärung wird in textueller Form gegeben. Diese ahmt den Menschen insofern nach, als dass der Mensch seine Entscheidungen in der Regel in vollständigen Sätzen begründet. Beispiele für textuelle Beschreibungen sind generierte Bildunterschriften oder Erklärverfahren, die mithilfe von Text, z. B. in Form von ganzen Sätzen, begründen, weshalb eine bestimmte Klasse vorhergesagt wurde. Beispielsweise könnte die Erklärung für die Klassifikation eines Bildes mit einem Wolf in natürlicher Sprache wie folgt lauten: *Das ist ein Wolf, weil er eine Schnauze hat.* Der erste Teil des Satzes *Das ist ein Wolf,* beschreibt die Entscheidung des Modells und der zweite Teil *weil er eine Schnauze hat* beinhaltet die Erklärung für die Entscheidung.

Grafiken: Die Erklärungen werden in Form einer Grafik dargestellt. Die Visualisierungen veranschaulichen, was ein Modell gelernt hat, indem bspw. die Parameter des Vorhersagemodells dargestellt werden. Ein Beispiel ist eine Heatmap, die aufzeigt, welche Bereiche für eine Entscheidung ausschlaggebend waren. Auf der nachstehenden Abb. 2.6 ist auf der rechten Seite eine Heatmap abgebildet, die die für die Klassifikation relevanten Bildbereiche hervorhebt. Auf der linken Seite der Abbildung ist das zu klassifizierende Originalbild dargestellt.

Abbildung 2.6: Heatmap zur Klassifizierung eines Wolfes [Quelle: [HHI21]].

Multimedia: Erklärbarkeit durch Multimedia kombiniert verschiedene Arten von Inhalten. Darunter fallen Texte, Grafiken, Animationen, Audios und Videos. Durch den Einsatz von Multimedia für eine Erklärung kann die Nachvollziehbarkeit – im Gegensatz zur bspw. nur textuellen Beschreibung – gesteigert werden.

2.3.2.3 Die Anwendergruppe

Es gibt verschiedene Anwendergruppen, für die Erklärungen generiert werden können. Je nach Anwendergruppe werden entsprechend den Erfahrungen und Zielen unterschiedliche Anforderungen an eine Erklärung gestellt [Wel17]. Erklärungen dienen unterschiedlichen Zwecken und können einen unterschiedlichen Grad an Komplexität aufweisen. Der Anwender kann z. B. ein *Endnutzer* ohne Fachwissen innerhalb der Domäne sein. Dieser benötigt einfache und leicht verständliche Erklärungen. Alternativ kann der Anwender ein *Domänenexperte* sein und mit den Eigenheiten der Daten, nicht aber mit den Verfahren vertraut sein. Einem solchen Anwender können komplexere Erklärungen präsentiert werden. Für einen Domänenexperten kann ein gewisser Grad an Komplexität gar zwingend notwendig sein, bspw. für einen

Arzt, der präzise und detaillierte Diagnosen vagen Beschreibungen vorzieht. Im Gegensatz dazu ist der *Entwickler* des Modells meist nicht mit der Domäne vertraut, verfügt jedoch über viel Erfahrung damit, sich mit komplexen technischen Konzepten auseinanderzusetzen. Eine Erklärung für einen Entwickler sollte daher technischer Natur sein und Interna eines Modells, die sich beispielsweise zur Inspektion (Debugging) eignen, enthalten. Ebenfalls von der Erfahrung des Anwenders abhängig ist der zeitliche Rahmen, der für das Verstehen der Erklärung benötigt wird. Eine Erklärung, die in kurzer Zeit verstanden werden muss, sollte in der Regel anders gestaltet sein als eine Erklärung, die einen längeren Zeitraum zum Verständnis erfordert.

2.3.3 Faktische und kontrafaktische Erklärungen

In diesem Abschnitt wird genauer zwischen faktischen und kontrafaktischen Erklärungen unterschieden, da diese für die Kapitel 4 und Kapitel 5 wichtig sind. Diese wurden bereits als Erklärarten in Abschnitt 2.3.2.1 eingeführt.

2.3.3.1 Faktische Erklärungen

Faktische Erklärungen beschreiben ein vorliegendes Ergebnis und erklären dieses anhand bestimmter verwendeter Merkmale. Die Mehrheit aktueller Erkläransätze, denen sich Kapitel 2.5 widmet, generiert faktische Erklärungen. Faktische Erklärungen können in unterschiedliche Arten eingeteilt werden und wurden bereits in Abschnitt 2.3.2.1 beschrieben. Diese können beispielhaft wie folgt ausgedrückt werden: Das Ergebnis für x ist y, weil $x = a$. Zum Beispiel lässt sich eine faktische Erklärung einfach als Entscheidungspfad aus einem Entscheidungsbaum extrahieren.

Faktische Erklärungen können im Gegensatz zu kontrafaktischen Erklärungen sowohl lokaler als auch globaler Natur sein. Ein konkretes Beispiel für eine lokale faktische Erklärung ist die Regel *Alter < 32, Beruf: Student, Einkommen < 500 Euro, KreditausfallRisiko = HohesRisiko*. Diese Regel drückt aus, dass eine Person, die jünger als 32 Jahre ist, vom Beruf Student ist und ein

Einkommen hat, das geringer als 500 Euro im Monat ist, ein hohes Kreditausfallrisiko birgt.

2.3.3.2 Kontrafaktische Erklärungen

Eine kontrafaktische Erklärung beschreibt die Änderung an den Merkmalswerten, die mindestens durchgeführt werden muss, um die ursprüngliche Vorhersage auf eine gewünschte Entscheidung zu verändern [Mol20]. Es werden also Hinweise darauf gegeben, wie eine gewünschte Entscheidung erzeugt werden kann [Wac17]. Kontrafaktische Erklärungen können beispielhaft wie folgt ausgedrückt werden: Wenn man statt x x' wählen würde, würde statt y die Entscheidung y' lauten. Bei einem potenziellen Kreditnehmer, der ein Einkommen in Höhe von 4000 Euro (x') anstatt 3000 Euro (x) erhält, wäre der Kreditantrag nicht abgelehnt, sondern genehmigt worden. Der kontrafaktische Aspekt beschreibt eine veränderte Realität mit $x' \rightarrow y'$, in der eine andere Entscheidung getroffen wird. Der Sachverhalt x hat die Konsequenz y. Ändert sich jedoch x in den kontrafaktischen Sachverhalt x', ist die Konsequenz daraus y'. Dabei sind x oder x' Eingabedaten in Form von Merkmalswerten für das Modell, und y bzw. y' sind die Ergebnisse.

Somit ist das Problem, eine kontrafaktische Instanz aufzufinden, ein Suchproblem im Merkmalsraum mit dem Ziel, den Punkt zu finden, der der gegebenen Instanz am nächsten liegt, jedoch zu einer anderen Klasse gehört. Eine gefundene kontrafaktische Instanz kann entweder für sich dargestellt werden (wie bei einer faktischen Erklärung) oder um gezielt diejenigen Veränderungen hervorzuheben, die für eine gewünschte Vorhersage ausschlaggebend sind.

Formal muss die folgende Formel 2.8 gelöst werden, um eine kontrafaktische Instanz zu finden:

$$\min_{x'} \max_{\lambda} \lambda(h(x') - y') + d(x, x') \tag{2.8}$$

wobei

$$d(x,x') = \sum_{i=1}^{N} \frac{|x_i - x'_i|}{\text{MAD}_i}, \tag{2.9}$$

wobei x' die kontrafaktische Instanz ist, y' die gewünschte Klasse der kontrafaktischen Instanz, $h(.)$ ist das Modell, $d(.)$ eine Metrik im Merkmalsraum und λ beschreibt einen Parameter, der die Klasse der kontrafaktischen Instanz mit dem Abstand zur faktischen Instanz gewichtet. N gibt die Dimension der Instanz und Mean Absolute Deviation (MAD) die mittlere absolute Abweichung des arithmetischen Mittels an. Die MAD wird wie folgt berechnet, wobei P ein Trainingsdatensatz ist:

$$MAD_i = median_{\vec{x} \in P} \left(x_i - \tilde{x}_i \right) \ . \tag{2.10}$$

Kontrafaktische Erklärungen sind (wie auch faktische Erklärungen) lokale Erklärungen, da diese eine Instanz erklären. Weiterhin handelt es sich bei kontrafaktischen Erklärungen um modell-agnostische Erklärungen, da diese nicht die internen Strukturen eines Modells verwenden müssen, um eine Lösung zu finden, sondern die Ein- und Ausgabedaten zur Generierung der Erklärung nutzen. Die Informationen, die zur Erstellung der Erklärung benötigt werden, sind die Eingabedaten und die jeweiligen Ergebnisse eines Modells. Globale kontrafaktische Erklärungen können nicht direkt gegeben werden. Es besteht die Möglichkeit, eine Art semi-globale kontrafaktische Erklärung zu erzeugen. Dabei kann durch eine repräsentative Anzahl an Instanzen eine globale Erklärung erzeugt werden wie bspw. durch das Verfahren *sp-lime* von Ribeiro et al. [Rib16a]. Passend zu den Instanzen können wiederum die kontrafaktischen Erklärungen ermittelt werden, um auf diese Weise eine Art globale kontrafaktische Erklärung zu erhalten.

In der Regel ist es möglich, mehrere passende kontrafaktische Erklärungen zu finden. Dementsprechend muss geprüft werden, welche Eigenschaften die

kontrafaktische Instanz genau erfüllen soll. Eine Forderung kann beispielsweise sein, dass nur bestimmte Merkmale zur Erreichung des gewünschten Ergebnisses geändert werden sollen. Zusätzlich muss in der Praxis darauf geachtet werden, nur die Merkmale zu ändern, die in der Realität auch abgeändert werden können. So ist es z. B. nicht sinnvoll, bei einem Kreditantrag den Wohnort zu ändern.

2.3.4 Ontologien zur Optimierung von Erklärungen

Ontologien können zur Optimierung von Erklärungen eingesetzt werden. Diese können dazu verwendet werden, um beispielsweise Erklärungen kompakter zu gestalten, indem bestimmte Merkmale zusammengefasst werden.

Wissen kombiniert Informationen, Erfahrungen und Fähigkeiten, um das Konzept des Verstehens zu ermöglichen. Die Grundlage dafür bilden Daten. Diese werden verwendet, um verschiedene Formen von Informationen zu erhalten und dabei wichtige Zusammenhänge zu erkennen. Ein einfacher Austausch von Wissen benötigt eine einheitliche Struktur und eine gemeinsame Definition. Dies ist vor allem wichtig, wenn Wissen zwischen Anwendungen ausgetauscht wird. Einen Ansatz hierzu liefern Ontologien. Gruber [Gru93] definierte erstmals eine Ontologie als Spezifikation einer Konzeptualisierung im Kontext des Wissensaustausches. Weiter beschreibt er die *Ontolingua* als eine Spezifikation einer Konzeptualisierung im Kontext des Wissensaustausches.

Diese Beschreibung umfasst Teile, die bis heute die wesentlichen Bausteine von Ontologien sind. Die Definitionen werden meist in einer standardisierten Sprache mit der Web Ontology Language (OWL) geschrieben. Die OWL[14] gilt als eine Sammlung von Sprachen, die zur Erstellung von Ontologien verwendet werden kann. Diese baut auf dem Ressource Description Framework auf, das ursprünglich vom World Wide Web Consortium (W3C)[15] zur Beschreibung von Metadaten konzipiert wurde. Dieses ist mittlerweile ein Kernstück

[14] https://www.w3.org/TR/owl-features/, Letzter Abruf am 05.01.2021.
[15] https://www.w3.org/, letzter Abruf am 05.01.2021.

des Semantic Webs[16]. Das Konzept definiert drei Tupelbeziehungen von Subjekten, Prädikaten und Objekten und kann als gerichteter Multigraph visualisiert werden. Die Darstellung von Wissen in Form von Ontologien bietet verschiedene Vorteile, wie bspw. die Ableitung von Informationen aus relationalen Konzepten. Daneben gibt es noch weitere Ansätze, die sich eignen, um relationales Wissen abzubilden. Dazu gehören z. B. *Taxonomien* oder *Wissensgraphen*.

Taxonomien [Kos93] werden meist verwendet, wenn Ursprünge und Verbindungen von Objekten aufgezeigt werden sollen. Diese werden dazu in Kategorien oder Klassen eingeteilt. Beispielsweise werden viele Wörterbücher in Form einer Taxonomie dargestellt. Auch Ontologien selbst verwenden in der Regel Taxonomien als Grundbaustein, um das Wissen besser strukturieren zu können.

Darüber hinaus können Ontologien selbst als Bausteine verwendet werden, wenn große Wissenszusammenhänge konzipiert werden sollen. Das Konzept wird als *Wissensgraph* [Ehr16] bezeichnet. Es entsteht eine High-Level-Struktur aus miteinander verbundenen Informationen. Diese werden verwendet, um ähnliche Themen miteinander zu verknüpfen, und eignen sich z. B. für Suchmaschinen. Einer der bislang größten Wissensgraphen wurde von Google erstellt. Dieser zeigt bei der Suche zusätzliche und verwandte Informationen an [Ede12] und soll einen schnellen Zugriff auf entsprechendes Wissen ermöglichen, indem ähnliche Informationen und häufig gesuchte Ergänzungen miteinander verknüpft werden.

Neben dem einheitlichen Austausch von Wissen über viele Bereiche hinweg werden Ontologien vor allem für das Erkennen von möglichen Zusammenhängen innerhalb einer Domäne verwendet. Im medizinischen Bereich kann die Suche nach den Ursachen einer Krankheit durch die Kategorisierung von identifizierten, expliziten Beziehungen innerhalb einer Ontologie unterstützen [Moh12]. Dies ermöglicht eine schnellere Verknüpfung von Symptomen und möglichen Ursachen. Ein weiteres Beispiel für den erfolgreichen Einsatz

[16] https://www.w3.org/standards/semanticweb/, letzter Abruf am 05.01.2021.

von Ontologien ist *Watson* von IBM [Hoy16]. Watson ist ein System zur Beantwortung von Fragen in natürlicher Sprache für unterschiedliche Domänen wie bspw. der bereits erwähnten Onkologie. Die Leistung des Systems basiert auf der hohen Menge an Trainingsdaten, die es verarbeiten kann, sowie der Fähigkeit, Beziehungen über Ontologien und ähnliche andere Konzepte zu verknüpfen. In vielen Anwendungsfällen wird mehr als eine Ontologie benötigt. Daher besteht die Notwendigkeit, Ontologien zu verbinden. Durch die Verfahren des maschinellen Lernens kann der Aufbau von großen Wissensbasen ermöglicht werden. Beispielsweise entwickelten Doan et al. [Doa04] ein System, das mithilfe von maschinellen Lernverfahren teilautomatisch semantische Mappings zwischen Ontologien erstellt.

Auch im Bereich der erklärbaren KI wird das Konzept der Ontologien bereits verwendet. Ontologien, die das Wissen über die verschiedenen Beziehungen von Daten repräsentieren, bergen ein großes Potenzial, um komplexe Datenstrukturen nachvollziehbarer zu gestalten. Publio et al. [Pub18] stellen eine Top-Level-Ontologie vor, die eine Reihe von Klassen, Eigenschaften und Einschränkungen enthält, die wiederum zur Definition von Informationen über maschinelle Lernverfahren, Datensätze und Experimenten verwendet werden können. Eine einheitliche Beschreibung dieser Elemente vereinfacht den Austausch des gesammelten Wissens. McGuinness et al. [McG07] beschreiben mit dem Namen *PML2* eine Kombination aus drei Ontologien, die als *Interlingua* für den Austausch von generierten Erklärungen geeignet sein soll. Diese besteht aus den Ontologien *Provenance*, *Justification* und *Trust-Relations*, die zur Beschreibung von Informationen innerhalb von Systemreaktionen verwendet und zur Generierung einer Erklärung genutzt werden können. Diese konzentriert sich auf die Informationen, wie ein System eine Ausgabe generiert hat, und deren Abhängigkeiten. PML2 soll eine einheitliche Definition schaffen, um die Bausteine einer Erklärung zu teilen.

Das Verfahren *Trepan Reloaded* [Con19] verwendet bei der Generierung der Erklärung ebenfalls eine Ontologie. Dabei wird ein Entscheidungsbaum als Surrogat verwendet, um den Entscheidungsprozess eines neuronalen Netzes zu erklären. Das Ziel ist es, das strukturierte Wissen innerhalb einer Ontologie zu nutzen und bei der Generierung des Entscheidungsbaums anzuwenden.

Die Nutzerstudie, die die Autoren durchführten, zeigte, dass Erklärungen, die der Struktur einer Ontologie folgen, die dem allgemeinen menschlichen Verständnis nach schlüssig sind, für die Anwender leichter zu verstehen waren. Panigutti et al. [Pan20] stellen in ihrer Arbeit *Doctor XAI* vor. Der Ansatz für medizinische Belange befasst sich mit sequenziellen, ontologisch verknüpften Daten. Das bedeutet, dass dieser Ansatz in der Lage ist, Verbindungen zwischen Informationen abzuleiten, die im Laufe der Zeit entstehen. Somit schafft dieser eine neue Wissensquelle, die zur Verbesserung der Vorhersagen und Erklärungen im medizinischen Bereich genutzt werden kann. Chen et al. [Che18] verwenden Ontologien, um Erklärungen für das Transferlernen zu implementieren. Das Ziel ist es, Erklärungen zu erstellen, die es Anwendern ohne technisches Vorwissen ermöglichen, fehlerhafte Übertragungen zu erkennen. Damit sollen die Übertragungen des Transferlernens optimiert werden. Geng et al. [Gen19] stellen einen alternativen Ansatz vor und verwenden Wissensgraphen. In diesem Zusammenhang betonen die Autoren das Potenzial von Wissensbasen für Erklärungen.

Lecue [Lec20] beschreibt die *Thales XAI Plattform*, in der Wissensgraphen eingebunden werden. Deren Ziel ist es, KI-Systeme für kritische Situationen anwendbar zu machen. Die Plattform soll Erklärungen in Form von beispielbasierten, merkmalsbasierten oder kontrafaktischen Erklärungen unter Verwendung textueller und visueller Darstellungen generieren. Zusätzlich werden Erklärungen auf Basis von Semantik mithilfe von Wissensgraphen generiert. Dabei werden semantische Hilfsmittel eingesetzt, um die Daten für das maschinelle Lernen mit Kontextinformationen anzureichern. Erklärungen werden abgerufen, indem ein Wissensgraph und die korrespondierenden Kontextinformationen der Ergebnisse der maschinellen Lernverfahren verwendet werden, um repräsentative semantische Beziehungen zu identifizieren.

2.3.5 Messen von Erklärbarkeit

Adadi et al. [Ada18] verweisen auf die Arbeiten von Breiman et al. [Bre01b], in denen beschrieben wird, dass Modelle des maschinellen Lernens oftmals

eine komplexe Struktur haben und für dieselben Eingabedaten der Algorithmus verschiedene Modelle mit gleicher Genauigkeit generiert. Dies kann damit erklärt werden, dass der Algorithmus unterschiedliche Pfade durchläuft. So kann es für identische Eingabedaten unterschiedliche Erklärungen geben, wodurch die Messung der Erklärbarkeit erschwert wird.

Doshi-Velez et al. [Dos17] erläutern eine Taxonomie zur Evaluierung für den Bereich der Erklärbarkeit. Sie unterscheiden zwischen anwendungsbezogener, nutzerbezogener und funktional-basierter Evaluierung. Beim anwendungs- und nutzerbezogenen Ansatz ist die Evaluation durch einen Anwender essenziell, um die Erklärbarkeit bewerten zu können. Die anwendungsbezogene Evaluierung beinhaltet die Durchführung von Versuchen in realen Anwendungen durch Domänenexperten. Dabei ist es vor allem von Interesse, herauszufinden, inwieweit das System für die angedachte Aufgabe funktioniert. Insbesondere die Qualität einer Erklärung wird im Zusammenhang mit der Erfüllung der Aufgabe gemessen. Das kann z. B. die Untersuchung der Fragestellung beinhalten, ob mithilfe der Erklärung eine bessere Identifizierung von Fehlern ausgeführt wird. Die nutzerbezogene Evaluierung versucht, einfache Experimente mit Anwendern ohne Fachkenntnisse durchzuführen, bei denen lediglich der Kern der Aufgabe erhalten bleibt. Diese Evaluierung eignet sich gut für anspruchsvolle Aufgaben. So kann zum Beispiel untersucht werden, welche Arten von Erklärungen am besten im Hinblick auf einen bestimmten Aspekt zu verstehen sind. Die funktional-basierte Evaluation benötigt keine Experimente mit Anwendern. Stattdessen werden Metriken zur Messung der Qualität einer Erklärung verwendet. Ein großer Vorteil des Ansatzes ist, dass dieser ressourcenschonend ist. Hierfür müssen keine aufwändigen Experimente vorbereitet und durchgeführt werden. Im Folgenden wird die Literatur im Bereich der Messung von Erklärbarkeit entsprechend der Taxonomie von Doshi-Velez et al. [Dos17] eingeordnet, da in Kapitel 6 Nutzerstudien durchgeführt werden, die in die nutzerbezogene Evaluierung eingeordnet werden.

2.3.5.1 Funktional-basierte Evaluierung

Insbesondere für regel- und baumbasierte Modelle gibt es verschiedene Maße, die zur funktional-basierten Evaluierung herangezogen werden können. So gibt es Maße, die die Gesamtzahl der verwendeten Merkmale verwenden [Su15]. Bei regelbasierten Modellen wird die Größe mit der Gesamtzahl der Regeln in einer Entscheidungsmenge gemessen, wie bspw. in den Arbeiten von Lakkaraju et al. [Lak16], Lakkaraju et al. [Lak17], Bertsimas et al. [Ber11] oder Letham et al. [Let15]. Die Länge einer Regel misst die Gesamtanzahl der Prädikate, die in der Bedingung des Entscheidungssatzes [Lak16] oder der Entscheidungsliste [Ber11] verwendet werden. Die Länge jeder einzelnen Regel kann kumuliert werden, um die Gesamtzahl der verwendeten Prädikate [Lak17] zu messen. Außerdem können Maße für die Gesamtzahl der Dateninstanzen, die eine Regel erfüllen (in [Lak16] als *Abdeckung* bezeichnet), und Maße für die Gesamtzahl der Dateninstanzen, die mehrere Regeln erfüllen (in [Lak16] als *Überlappung* bezeichnet), zur Messung der Komplexität verwendet werden.

Freitas [Fre14] schlägt ein anderes Maß für die Komplexität eines regelbasierten Modells vor. Er argumentiert, dass die durchschnittliche Anzahl von Regelbedingungen, die für die Erstellung von Vorhersagen berücksichtigt wurden, ein besseres Komplexitätsmaß darstellt. Für Entscheidungsbäume kann die gesamte Anzahl der Merkmale, die als Aufteilungsmerkmale gelten, verwendet werden [Cra96]. Ein weiteres Maß für die Komplexität ist die *Tiefe* [Rüp05]. Weiterhin kann die Merkmalsbedeutung aus regelbasierten und baumbasierten Modellen extrahiert werden. Die Gesamtzahl der Instanzen, die ein Merkmal für die Klassifikation verwenden, kann als die Wichtigkeit des Merkmals verwendet werden. Samek et al. [Sam17] bedienen sich der Perturbationsanalyse zur Messung der Erklärbarkeit. Diese Methode basiert auf drei einfachen Ideen. Zunächst ist die Annahme, dass der Vorhersagewert eines Modells anfälliger gegenüber wichtigen als gegenüber unwichtigen Merkmalen ist. Weiter generieren Ansätze wie die *Sensitivitätsanalyse* oder die *Layer Wise Relevance Propagation* einen Merkmalswert, der es ermöglicht, diese zu sortieren. Abschließend ist es möglich, die Eingabewerte iterativ zu ändern

und dabei die Vorhersagewerte zu dokumentieren. Der gemittelte Vorhersagewert kann zur Messung der Erklärqualität verwendet werden. Schwankt der gemittelte Vorhersagewert stark, kann dies ein Hinweis auf wichtige oder unwichtige Parameter einer Erklärung sein.

Verfahren im Bereich der Post-hoc-Erklärbarkeit, die auf einem Modell wie einer Blackbox basieren, können anhand einer Metrik im Vergleich zum Ursprungsmodell bewertet werden. Das Maß der Modelltreue gibt Aufschluss darüber, wie gut ein Surrogat die Vorhersagen des Originalmodells nachahmt. Die Modelltreue wird auch mithilfe der Metriken Genauigkeit, F1-Score, des Recalls oder der Präzision berechnet. Diese werden in Bezug auf die Ergebnisse der Blackbox gemessen. Die Modelltreue kann sowohl auf modell-spezifische, modell-agnostische als auch auf lokale oder globale Verfahren angewendet werden. Die nachfolgende Gleichung (2.11) zeigt die Berechnung der Modelltreue

$$\text{Modelltreue} = 1 - \mathcal{P}\big(w(x) \neq b(x), \mathcal{D}_{\mathcal{I}}\big), \tag{2.11}$$

wobei w das interpretierbare Modell und b das Blackbox-Modell ist, die die Ergebnisse auf einem Testdatensatz $\mathcal{D}_{\mathcal{I}}$ generieren. Ante-hoc-Verfahren können nicht durch die Modelltreue bewertet werden, da hierbei kein weiteres Modell generiert wird. Molnar et al. [Mol20] nennt jedoch noch weitere Metriken wie die Stabilität, die Konsistenz, den Grad der Wichtigkeit, die Repräsentativität oder die Verständlichkeit. Er beschreibt diese in Verbindung mit den folgenden Fragestellungen:

- Stabilität: Wie ähnlich sind sich Erklärungen für ähnliche Instanzen?
- Konsistenz: Wie stark unterscheidet sich eine Erklärung zwischen Modellen, die auf die gleiche Aufgabe trainiert wurden?
- Grad der Wichtigkeit: Wie gut spiegelt die Erklärung die Bedeutung von Merkmalen wider?
- Representativität: Wie viele Instanzen umfasst eine Erklärung?

- Verständlichkeit: Wie gut verstehen Anwender die Erklärungen? [Mol20]

Der Punkt der Verständlichkeit ist für die anwendungs- und die nutzerbezogene Evaluierung, die im nächsten Abschnitt beschrieben wird, essenziell da zwingend Anwender hierfür miteinbezogen werden.

2.3.5.2 Anwendungs- und nutzerbezogene Evaluierung

Die anwendungs- und die nutzerbezogene Evaluierung wurden zusammengefasst, da beide aktiv Anwender in Form von Benutzerstudien in die Evaluierung einbinden. Benutzerstudien im Bereich der Erklärbarkeit bilden einen wichtigen Forschungsschwerpunkt. Nur indem zusammen mit Anwendern untersucht wird, inwieweit Erklärungen beim Verstehen der Modelle bzw. der Modellergebnisse unterstützen, können die Vorteile von Erklärungen aufgezeigt werden. In der Regel wird ein vorab definiertes Maß benötigt, um die Nachvollziehbarkeit messen zu können.

Ist für eine bestimmte Aufgabe bereits Domänenwissen vorhanden, wird dem Ergebnis eines Vorhersagemodells mehr vertraut, wenn die Entscheidung in Übereinstimmung mit dem bereits vorhandenen Domänenwissen steht. Auf Basis dieser Grundlage kann das Vertrauen als das Verhältnis der Anzahl der akzeptierten Entscheidungen und der Anzahl aller Entscheidungen gemessen werden. Poursabzi-Sangdeh et al. [Pou21] messen Vertrauen, indem diese den Unterschied zwischen der Vorhersage des Modells und der Vorhersage der Versuchsteilnehmer bestimmen. Die Aufgabe der Versuchsteilnehmer war es, Kaufpreise von Häusern vorherzusagen. Die Versuchsteilnehmer erhielten Informationen über das zugrundeliegende Vorhersagemodell und sollten selbst

eine Vorhersage für den Kaufpreis des Hauses treffen. Diese Schätzung wurde mit dem vom Modell vorhergesagten Preis des Hauses verglichen. Die absolute Abweichung wurde als Maß für das Vertrauen herangezogen. Kleinere Werte spiegelten ein höheres Vertrauen und größere Werte ein niedrigeres Vertrauen in das Modell wider. Auch Ribeiro et al. [Rib16b] führten eine Nutzerstudie durch, um herauszufinden, ob Versuchsteilnehmer den Ergebnissen des Vorhersagemodells vertrauen. Zunächst erhielten die Teilnehmer zehn verschiedene Beispiele zu Häusern einschließlich der dazugehörigen Vorhersagen. Acht von zehn dieser Vorhersagen wurden von der Blackbox korrekt prognostiziert, während zwei Objekte falsch klassifiziert wurden. Die Teilnehmer wurden gebeten, die folgenden drei Fragen zu beantworten: „Vertrauen Sie dem Algorithmus, sodass dieser in der realen Welt gut funktionieren würde? Warum? Warum glauben Sie, dass der Algorithmus in der Lage ist, zwischen den Klassen zu unterscheiden?“ Als nächstes erhielten die Versuchsteilnehmer zehn verschiedene Beispiele zusammen mit den dazugehörigen Erklärungen und wurden erneut mit den denselben Fragen konfrontiert. Die beiden Versuchsausprägungen wurden anschließend gegeneinander ausgewertet, um festzustellen, ob die Bereitstellung einer Erklärung die Nachvollziehbarkeit erhöht und ob Erklärungen das Vertrauen beeinflussen. Die Untersuchungen zeigten, dass die Bereitstellung einer Erklärung dazu beiträgt, die Nachvollziehbarkeit eines Modells zu steigern.

Lage et al. [Lag19] führten eine weitere Nutzerstudie durch, um herauszufinden, was Erklärungen für den Menschen nachvollziehbar macht. Dazu wurden die Eigenschaften einer Erklärung systematisch verändert, um ihre Wirkung im Hinblick auf die Erfüllung verschiedener Aufgaben zu messen. Die Aufgaben bestanden darin, die Reaktion des Systems zu simulieren, eine vorgeschlagene Vorhersage zu verifizieren und kontrafaktische Erklärungen zu prüfen. Die Studie zeigte auf, dass kontrafaktische Erklärungen über die gesamten Experimente hinweg eine signifikant geringere Genauigkeit beim Lösen der Aufgabe aufwiesen. Auch wurde eine längere Reaktionszeit bei kontrafaktischen Erklärungen gemessen. Schmidt et al. [Sch19b] führten ein quantitatives Maß für das Vertrauen in Entscheidungen ein. In dem Experiment wurden zwei Methoden untersucht: *COVAR*, mit dem ein direkt interpretierbares Modell erzeugt wird, und *LIME*, das eine Erklärung

basierend auf einer Blackbox mithilfe eines lokalen Surrogates (lineares Modell) erzeugt. Das Ergebnis der Studie war, dass die COVAR-Methode nachvollziehbare Erklärungen lieferte. Dabei wurde vor allem der Vorteil der Verwendung einfacher, interpretierbarer Modelle hervorgehoben. Read et al. [Rea93] untersuchten die Hypothese, dass Menschen simple und generalisierte Erklärungen stärker bevorzugen. Die Versuchsteilnehmer sollten die Wahrscheinlichkeit und die Qualität von Erklärungen anhand unterschiedlicher Ursachen bewerten. Diese wurden mit Szenarien konfrontiert, die mehrere zu erklärende Ereignisse und Erklärungen enthielten. Beispielsweise handelte ein Szenario von einer Frau, die an Gewichtszunahme, Müdigkeit und Übelkeit litt. Die Teilnehmer erhielten Erklärungen, die für die möglichen Symptome bei der Frau ausschlaggebend waren: Eine Erklärung war, dass die Frau aufgehört hatte zu trainieren (Grund für die Gewichtszunahme). Die zweite Erklärung besagte, dass die Frau an Mononukleose litt (Grund für die Müdigkeit) und einen Magenvirus hatte (Grund für die Übelkeit). Als dritte Erklärung hieß es, dass die Frau schwanger war (Grund für alle drei Symptome). Die Teilnehmer bevorzugten alle die einfachste Erklärung, mit der sich alle drei Symptome begründen ließen, und vermuteten eine Schwangerschaft.

Zusammenfassend zeigen die Ergebnisse der Experimente, dass vor allem einfache Erklärungen, die mit der Erfahrung des Anwenders übereinstimmen, bevorzugt werden. Die Literaturrecherche zeigt jedoch auch, dass es kein einheitliches Vorgehen bei der Messung der Erklärbarkeit im Hinblick auf Nutzerstudien gibt. Jede der Nutzerstudien stellt eine eigene Metrik zur Messung der Erklärbarkeit auf, was den Vergleich der Ergebnisse der Nutzerstudien erschwert. Dies ist ebenso durch die unterschiedlichen Aufgabenstellungen bedingt, d. h., ob es sich dabei um eine reine Verifikation der Vorhersage handelt, wie z. B. bei Ribeiro et al. [Rib16a] oder Poursabzi-Sangdeh et al. [Pou21], oder ob eine Simulation des Modells vorgenommen wird, wie bei Schmidt et al. [Sch19b].

2.3.5.3 Quantitative Auswertungsmethoden bei der anwendungs- und nutzerbezogenen Evaluierung

Hypothesentests überprüfen Annahmen über Messwerte mit statistischen Methoden. Die Nullhypothese wird dazu formuliert um zu testen ob diese verworfen oder doch beibehalten wird. Die Untersuchungen sollen aufzeigen, ob gemessene Werte unter einer getroffenen Annahme unwahrscheinlich sind. Sofern die Messwerte der getroffenen Annahme in einem bestimmten Wahrscheinlichkeitsbereich liegt, wird die Annahme beibehalten. Zur Feststellung signifikanter Zusammenhänge werden beispielsweise p-Werte verwendet [Kur20]. Der p-Wert[17] liegt zwischen dem Bereich Null bis Eins, da dieser eine Wahrscheinlichkeit angibt. Fröhlich et al. [Frö09] merken an, dass statistische Signifikanz lediglich etwas über die Existenz eines Effektes, nicht jedoch über dessen Bedeutsamkeit und Relevanz im Kontext der Fragestellung aussagt. Somit ist es notwendig, auch die Relevanz von signifikanten Ergebnissen zu untersuchen. Hierfür gibt es in der Literatur unterschiedliche Effektmaße [Rus03]. Beispielsweise kann dazu *Cohens d* verwendet werden. Hierbei handelt es sich um ein Effektmaß für die Untersuchung der Unterschiede des Mittelwerts zwischen zwei Gruppen [Kup11].

2.4 Regularisierung

Zur Verbesserung der Genauigkeit auf den Testdaten von Modellen kann das Konzept der Regularisierung eingesetzt werden. Die Regularisierung kann jedoch auch dazu verwendet werden, um die Erklärbarkeit eines Modells zu unterstützen. In diesem Abschnitt werden zunächst die Grundlagen der Regularisierung und bekannte Verfahren beschrieben, durch die ein Modell regularisiert werden kann. Darauf aufbauend werden Verfahren vorgestellt, die

[17] Der p-Wert gibt die Wahrscheinlichkeit über die Evidenz der Nullhypothese an. Kleine p-Werte werden so interpretiert, dass diese eine starke Evidenz aufzeigen [Du 09]. Ist der p-Wert kleiner als ein vorgegebenes Signifikanzniveau, lässt sich die Nullhypothese ablehnen. Wird die Nullhypothese anstelle der Alternativhypothese verworfen, bezeichnet man das Ergebnis als signifikant.

die Regularisierung gezielt zur Verbesserung der Nachvollziehbarkeit verwenden.

2.4.1 Regularisierung im Allgemeinen

Die Regularisierung ist ein wichtiges Mittel des maschinellen Lernens, das es erlaubt, Modelle mit einer hohen Genauigkeit zu trainieren, die gut generalisieren. Insbesondere kann eine *Überanpassung* der Modelle verhindert werden. Überanpassung bedeutet, dass sich das Modell zu stark an den Trainingsdaten orientiert, wodurch es nur noch begrenzt generalisiert. Bei der Unteranpassung vergrößert sich der Fehler des Modells, da die Zusammenhänge innerhalb der Daten nicht ausreichend berücksichtigt werden. Die Regularisierung beschreibt im Allgemeinen die Technik, bei der die Zielfunktion eines Modells einen Strafterm hinzugefügt bekommt, um gewünschte Eigenschaften zu erreichen. Das kann z. B. die Reduzierung der Komplexität eines Modells sein, um die Nachvollziehbarkeit des Modells gewährleisten zu können. Goodfellow et al. [Goo16] beschreiben die Regularisierung als „jede Änderung, die an einem Lernalgorithmus vorgenommen wird, um den Testfehler aber nicht den Trainingsfehler zu reduzieren“[18].

Die klassische Verlustfunktion aus Gleichung (2.1) wird dabei um den zweiten Term erweitert. Daraus resultiert die Gleichung (2.12).

$$L'(\theta, \mathcal{D}) = L(\theta, \mathcal{D}) + \lambda \cdot \Omega(\theta) \tag{2.12}$$

Der erste Term der Gleichung stellt den Fehler auf den Daten und der zweite Term den Strafterm der Modelle dar. λ spiegelt die Stärke der Bestrafung wider, Ω die Komplexität des Modells, θ die Parameter des Modells und D den Datensatz. Bei der Minimierung der Funktion wird somit versucht, den

[18] [Goo16], S. 117, Eigene Übersetzung: „any modification we make to a learning algorithm that is intended to reduce its test error but not its training error“

Trainingsfehler zu verringern und die Generierung komplexer Modelle zu bestrafen. Bei der Wahl von λ muss darauf geachtet werden, den Wert nicht allzu groß zu wählen, da anderenfalls zu einfache Modelle generiert werden, die starke Verzerrungen erlauben können [Alp19]. Nachfolgend werden die bekanntesten Arten der Regularisierung beschrieben.

L1-Norm: Die L1-Norm versucht, den Koeffizienten unbedeutender Merkmale auf Null zu setzen. Dabei wird die absolute Größe der Koeffizienten verwendet (s. Gleichung (2.13)).

$$\Omega(\theta) = \sum_i \sum_j |\theta_{ij}| \tag{2.13}$$

Dies führt dazu, dass die Werte stark eingeschränkt und einige der geschätzten Parameter zu Null werden. Die Gleichung ist in Gleichung (2.14) dargestellt. Je stärker die Bestrafung ist, desto mehr verkleinern sich auch die Schätzungen der Parameter gegen Null. Daraus resultieren weniger Merkmale im Modell.

$$L'(\theta, \mathcal{D}) = L(\theta, \mathcal{D}) + \lambda \cdot \|\theta\|_1 \tag{2.14}$$

L2-Norm: Bei der L2-Norm (oder Ridge-Regression) besteht der Regularisierungsterm aus der Summe der Quadrate aller Merkmalsgewichte. Dies ist in Gleichung (2.15) dargestellt. Die L2-Norm führt zu kleinen Werten, die im Gegensatz zur L1-Norm nicht viele Nullwerte enthalten. Daraus resultiert in der Regel eine nicht dünn besetzte Lösung. Die L2-Regularisierung ist gegenüber Ausreißern anfällig, da der quadratische Term die Fehlerdifferenzen der

Ausreißer vergrößert. Der Regularisierungsterm versucht, dies durch die Bestrafung der Gewichte zu beheben. Die Verlustfunktion mit der L2-Norm ist in Gleichung (2.16) dargestellt.

$$\Omega(\theta) = \sum_i \sum_j {\theta_{ij}}^2 \tag{2.15}$$

$$L'(\theta, \mathcal{D}) = L(\theta, \mathcal{D}) + \lambda \cdot ||\theta||_2^2 \tag{2.16}$$

Elastisches Netz: Die Kombination aus L1- und L2-Norm wird in der Literatur als *elastisches Netz* bezeichnet [Zou05]. Daraus ergibt sich die Gleichung (2.17).

$$L'(\theta, \mathcal{D}) = L(\theta, \mathcal{D}) + \lambda \cdot ||\theta||_1 + ||\theta||_2^2 \tag{2.17}$$

Dropout *Dropout* beschreibt eine Technik bei der während dem Training, Einheiten (zusammen mit ihren Verbindungen) randomisiert aus dem neuronalen Netz gelöscht werden. Während des Trainings werden durch das Dropout mit einer bestimmten Wahrscheinlichkeit (z. B. bei neuronalen Netzen) bestimmte Neuronen ausgelassen. Dabei ist der gewünschte Effekt, dass das Netz weniger empfindlich gegenüber spezifischen Gewichten der Neuronen wird. Dies wiederum führt zu einem Netz, das in der Regel besser generalisiert und weniger zu einer Überanpassung der Trainingsdaten neigt.

2.4.2 Regularisierung im Bereich des erklärbaren maschinellen Lernens

Die Regularisierung wird auch im Bereich des erklärbaren maschinellen Lernens eingesetzt, um die Nachvollziehbarkeit der Modelle zu unterstützen. In

der Praxis garantiert ein vermeintlich einfaches Modell, wie bspw. ein lineares Regressionsmodell, per se nicht die Erklärbarkeit. Ein lineares Modell im hochdimensionalen Bereich kann leicht nicht nachvollziehbar werden. Um dieses Problem zu vermeiden, kann die Anzahl der Eingabemerkmale während des Modelltrainings durch die Regularisierung begrenzt werden, in dem unbedeutende Eingabemerkmale vernachlässigt werden. Dazu kann beispielsweise die L1- oder die L2-Norm verwendet werden.

Hu et al. [Hu16] demonstrieren, wie ein neuronales Netz regularisiert wird, indem es sich an vordefinierte Logikregeln hält und dabei dem neuronalen Netz Vorwissen über die Domäne vermittelt wird. Die Autoren verwenden dazu die Posteriori-Regularisierung nach Ganchev et al. [Gan10]. Bei dieser wird die Inferenz über die Posteriori-Wahrscheinlichkeit unter Berücksichtigung bestimmter Nebenbedingungen beschrieben. Hinton et al. [Hin15] regularisieren, um ein kleineres Modell mit den gleichen Entscheidungsgrenzen zu erhalten. Dazu versuchen sie, die Generalisierungsfähigkeit des komplexeren Modells auf ein kleines Modell zu übertragen. Die Klassenwahrscheinlichkeiten, die vom komplexen Modell erzeugt wurden, werden für das Training des kleineren Modells als „weiche Ziele“ verwendet. Ist das komplexe Modell ein Ensemble von einfacheren Modellen, kann das arithmetische oder geometrische Mittel der individuellen Vorhersageverteilungen als weiches Ziel verwendet werden. Wu et al. [Wu18] verwenden die Regularisierung, um Modelle zu generieren, die durch ein globales Surrogat-Modell in Form eines Entscheidungsbaumes nachgeahmt werden können. Die Regularisierung wird auf Gated Recurrent Units (GRUs) angewendet, die für die Zeitreihenvorhersage verwendet werden. Die Surrogat-Entscheidungsbäume werden während des Trainings der GRUs wiederholt aufgebaut. Aus den Entscheidungsbäumen wird die durchschnittliche Entscheidungspfadlänge zur Quantifizierung der Modellgröße abgeleitet. Diese Metrik wird als Regularisierungsterm in das Training zurückgeführt. Die Herausforderungen des Vorgehens zur Regularisierung von Wu et al. [Wu18] bestehen darin, dass zum einen die Laufzeiten zum Training der Modelle schnell steigen können, und zum anderen, dass eine sorgfältige Bestimmung der Parameter erforderlich wird. In einer weiteren Arbeit von Wu et al. [Wu19] wird ein Verfahren vorgestellt, das ein Blackbox-Modell durch mehrere separate Entscheidungsbäume, die für vordefinierte

Regionen gültig sind, approximiert. Die Regularisierung zielt darauf ab, nur die durchschnittliche Entscheidungspfadlänge der komplexesten Regionen zu bestrafen. Die Idee von Schaaf et al. [Sch19a] ist es, die Orthogonalität der Gewichtsvektoren zu berücksichtigen, um das neuronale Netz zu trainieren. Dabei wird ein globales Surrogat in Form eines Entscheidungsbaums erzeugt, der zur besseren Nachvollziehbarkeit der Ergebnisse verwendet werden soll. Im Gegensatz zu den Arbeiten von Wu et al. [Wu18] wird das Surrogat nicht wiederholt aufgebaut, was sich positiv auf die Laufzeit niederschlägt.

Ross et al. [Ros17] legen in ihrer Arbeit dar, wie lokale Erklärungen mit Domänenwissen während des Trainings eingeschränkt werden. Es wird davon ausgegangen, dass das Domänenwissen binär in einer Matrix kodiert ist. Diese gibt an, ob ein Merkmal verwendet werden soll, um eine Vorhersage zu treffen. Die Verlustfunktion, die zum Trainieren des Modells verwendet wird, wird sodann um einen neuen Term erweitert. Dieser dient dazu, Eingabegradienten zu bestrafen, die nicht mit dieser Annotationsmatrix übereinstimmen. In der Evaluation zeigen die Autoren, dass Modelle mit dem vorgeschlagenen Ansatz deutlich besser generalisieren.

2.5 Grundlegende Verfahren der Erklärbarkeit

Im Mittelpunkt dieses Kapitels stehen grundlegende Verfahren zur Erzeugung von Erklärungen. Die Verfahren der Klassifikation wurden in Abschnitt 2.1.2.1 zwischen Blackbox- und Whitebox-Modellen unterschieden. In Abschnitt 2.1.2.1 wurden vorwiegend Blackbox-Modelle erläutert. Nachfolgend werden die Whitebox-Modelle, die die Grundlage vieler Erklärverfahren bilden, beschrieben. Zu jedem Konzept wird beispielhaft ein Erklärverfahren, das dieses verwendet, vorgestellt. Diese bilden die Grundlage für die in Kapitel 3 vorgestellten Verfahren zur Erzeugung von Erklärungen. Um die Vorgehensweise einzelner Verfahren besser darzustellen, wird der IRIS-Datensatz [Fis36] verwendet. Dieser enthält 150 Beobachtungen mit vier Attributen unterschiedlicher Arten von Schwertlilien. Die Arten sind Iris setosa, Iris virginica oder Iris versicolor.

2.5.1 Lineare Modelle

Lineare Modelle zur Klassifikation werden durch die logistische Regression erzeugt. Jedem Merkmal wird ein Gewicht zugeordnet. Dieses zugewiesene Gewicht gibt den Beitrag des Merkmals zur Vorhersage an. Das Ergebnis der logistischen Regression gibt eine Wahrscheinlichkeit zwischen den Zahlen Null und Eins zurück. Die Gewichte beeinflussen die Wahrscheinlichkeit nicht mehr linear. Die gewichtete Summe wird durch die logistische Funktion in eine Wahrscheinlichkeit transformiert [Mol20].

Das Verfahren Local Interpretable Model-Agnostic Explanations (LIME) [Rib16a] ist eines der bekanntesten modell-agnostischen Post-hoc-Ansätze, das ein lineares Modell als Surrogat erzeugt. Dort wird, nachdem die Ergebnisse eines Blackbox-Verfahrens zur Verfügung stehen, in der lokalen Nachbarschaft der vorliegenden Instanz ein Surrogat (z. B. lineares Modell oder Entscheidungsbaum) trainiert. Darauf basierend werden für eine bestimmte Instanz die für die Vorhersage ausschlaggebendsten Merkmale (grün:positiv und orange:negativ) graphisch dargestellt. Die Abb. 2.7 zeigt die grafische Darstellung von LIME. Ausgehend von einer vorliegenden Instanz werden in der Nachbarschaft der Instanz ähnliche Instanzen über ein zufälliges Sampling generiert. Anhand dieser gesampelten Instanzen wird z. B. ein einfaches lineares Modell gelernt, aus dem die Merkmalsrelevanz extrahiert werden kann.

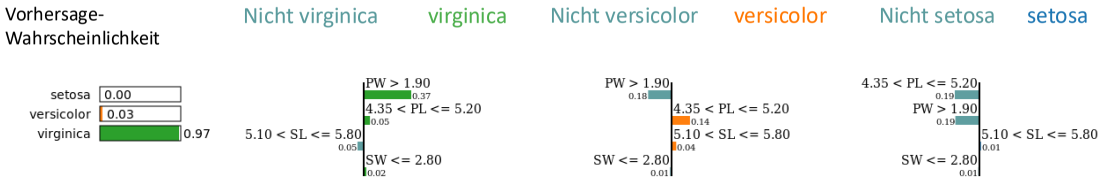


Abbildung 2.7: Grafische Darstellung von LIME [Quelle: In Anlehnung an [Bur21b], S. 279].

Eine spezielle Form von linearen Modellen sind Scoring-Systeme. Diese weisen jedem Merkmal oder Merkmalsintervall ein Gewicht zu. Ein endgültiger *Score* wird auf die gleiche Weise wie die Vorhersage eines linearen Modells bewertet. Die Klassifizierung einer Instanz erfolgt nach dem Vergleich des endgültigen Scores basierend auf einem definierten Schwellwert. Als Ansatz, um Scoring-Systeme nachvollziehbarer zu gestalten, eignet sich SLIM [Ust14]. Dabei wird ein pareto-optimaler Kompromiss zwischen der Genauigkeit und der Spärlichkeit der Parameter eingeführt.

Generalisierte Additive Modelle (GAMs) lernen eine lineare Kombination von Formfunktionen, um die Beziehung jedes Merkmals mit dem Ziel zu visualisieren [Lou12]. Eine Formfunktion setzt ein einzelnes Merkmal mit dem Ziel in Beziehung. Die Visualisierung einer Formfunktion veranschaulicht daher den Beitrag des Merkmals zur Vorhersage. Formfunktionen können Regressionssplines, Bäume oder Ensembles von Bäumen sein. Interaktionen zwischen den Merkmalen werden von GAMs jedoch nicht berücksichtigt. Lou et al. [Lou13] führen den Interaktionsbegriff in GAMs ein. Das Generalized Additive Models plus Interactions (GA^2Ms) berücksichtigt paarweise Interaktionen von Merkmalen. Diese Interaktionsterme erhöhen die Genauigkeit des Vorhersagemodells und erhalten gleichzeitig die Interpretierbarkeit des Vorhersagemodells.

2.5.2 Logische Aussagen

Eine logische Aussage ist ein deklarativer Satz, der als wahr oder falsch ausgewertet werden kann. Solche Aussagen können in einzelnen Entscheidungsregeln, Regellisten, Entscheidungsbäumen oder Entscheidungstabellen kombiniert werden. Sofern diese Kombinationen von Aussagen einfach gehalten werden, werden diese als nachvollziehbar bezeichnet. Im Folgenden werden die genannten Ansätze näher erläutert da bspw. Regellisten im weiteren Verlauf der Arbeit verwendet werden.

Entscheidungsregeln: Entscheidungsregeln haben die Form *WENN (engl. IF)* Bedingung *DANN (engl. ELSE)* Label. Um eine weitere Regel zu einer bestehenden Liste hinzuzufügen, kann diese durch *SONST (ELSE IF)* erweitert werden. Eine Bedingung in einer Regel besteht entweder aus einem einzelnen Merkmal, Operator, Wert-Tripel (in der Literatur manchmal auch als Literal bezeichnet) [Wan15c] oder einem Prädikat [Lak17, Lak16] (oder aus einer Konjunktion oder Disjunktion von Prädikaten). Der häufigste Fall ist die Verwendung einer Konjunktion von Prädikaten als Bedingung [Huy11]. Eine Regel, die nur eine Bedingung verwendet, wird als einstufige Regel bezeichnet. Entscheidungsregeln haben eine textuelle Darstellung. Diese sind entweder propositional oder schräg. Propositionale Entscheidungsregeln verwenden nur Merkmale als Prädikate von Bedingungen. Werden mehrere Prädikate als Bedingung verwendet, können diese durch die Disjunktion oder die Konjunktion kombiniert werden. Schräge Entscheidungsregeln verwenden lineare Kombinationen von Merkmalen als Bedingungen. Außerdem können die in einer Entscheidungsregel verwendeten Bedingungen als eine Beschreibung für die Klasse angesehen werden, die die Bedingung vorhersagt [Wan15c].

Regellisten: Regellisten sind geordnete Entscheidungsregeln, die eine neue Instanz basierend auf der ersten übereinstimmenden Regel klassifizieren [Fre14]. Die Listen werden normalerweise konstruiert, indem zuerst Entscheidungsregeln gefunden und anschließend geordnet werden [Let12]. Um eine neue Instanz zu klassifizieren, muss am Anfang der Liste begonnen und die von der Bedingung der Regel verwendeten Merkmale müssen mit denen der Instanz verglichen werden. Sofern es keine vollständige Übereinstimmung gibt, muss die nächste Regel in der Liste geprüft werden. Stimmen die Merkmale der Instanz mit keiner der Bedingungen der Regeln in der Liste überein, wird die Instanz durch eine Standardregel klassifiziert. Da eine Regel in einer Entscheidungsliste nur im Kontext der vorherigen Regeln gültig ist, steigt die Komplexität von Entscheidungslisten mit jeder zusätzlichen Regel stark an. In Listing 2.1 ist ein Beispiel einer Regelliste dargestellt, die die Pflanzenart *Setosa* klassifiziert.

```
WENN Blütenblattbreite (cm) : 0.8 bis unendlich DANN
    Wahrscheinlichkeit für die Klasse Setosa: 1.2 %
WENN SONST Blütenblattlänge (cm) : - unendlich bis 2.45 DANN
    Wahrscheinlichkeit für die Klasse Setosa: 97.4 %
SONST Wahrscheinlichkeit für die Klasse Setosa: 50.0 %
```

Listing 2.1: Beispiel einer Regelliste für die Pflanzenart *Setosa* erzeugt durch das Verfahren *BRL* [Quelle: In Anlehnung an [Bur21b], S. 25].

Scalable Bayesian Rule List (SBRL) [Yan17] ist eine Optimierung des Algorithmus Bayesian Rule List (BRL) [Let15]. Das Verfahren erzeugt eine geordnete Liste von Regeln, indem häufige Muster im Datensatz aufgefunden werden. Diese sind die Antezedenten der Regeln und können beispielsweise mit dem Algorithmus *Apriori* [Agr94] oder *Eclat* [Zak97] aufgebaut werden. Der *Support* eines Musters wird als die Menge aller Datenpunkte beschrieben, die dieses Muster erfüllen. Eine Menge von Bedingungen wird nicht mehr ergänzt, wenn das Hinzufügen weiterer Bedingungen den Support unter eine bestimmte Grenze fallen lässt. Nachdem die Muster gefunden wurden, wird die eigentliche Regelliste gebildet. Dabei geht der Algorithmus von einer Verteilung der möglichen Listen aus. Um die Interpretierbarkeit zu erreichen, macht er einige Annahmen über die Verteilung von Eigenschaften der Liste wie z. B. der Länge. Anschließend wird eine Liste aus dieser Verteilung gezogen, die iterativ modifiziert wird. Der Algorithmus gibt die Liste mit der höchsten A-posteriori-Wahrscheinlichkeit zurück.

Der CN2-Algorithmus [Cla89] erzeugt eine ungeordnete Regelliste. Der Algorithmus ist in verschiedenen Implementierungen verfügbar wie z. B. in der von Demšar et al. [Dem13], die auf der Arbeit von Clark et al. [Cla91] basiert. Es wird eine Suche durchgeführt, um mögliche Listen von Antezedenzen zu suchen. Der Suchbaum wird im Anschluss optimiert und verkürzt.

Der Algorithmus Repeated Incremental Pruning to Produce Error Reduction (RIPPER) wurde von Cohen [Coh95] als eine Verbesserung des IREP-Algorithmus [Für94] eingeführt. Er besteht aus zwei Phasen: Einer

Wachstumsphase, die Regeln zu einer Menge hinzufügt, und einer Optimierungsphase, die inkrementell Varianten der Regeln erzeugt und stutzt. Das Wachstum in beiden Phasen wird durch den Informationsgewinn gemäß der Definition von Quinlan [Qui90] gesteuert.

Entscheidungsbaum: Ein Entscheidungsbaum besteht sowohl aus Baum- als auch aus Blattknoten. Dem Baumknoten wird ein Aufteilungsmerkmal und ein Aufteilungswert zugewiesen. Dem Blattknoten wird eine Zielvariable zugeordnet. Der Prozess der Klassifizierung beginnt am obersten Baumknoten, der auch *Wurzelknoten* genannt wird. Basierend auf dem Aufteilungskriterium (bestehend aus Aufteilungsmerkmal und Aufteilungswert) werden die Pfade zu einem Blattknoten gewählt, um eine Zielvariable zu erhalten. Entscheidungsbäume werden von oben nach unten konstruiert, sodass ein einmal ausgewähltes Merkmal sowie ein Merkmalswert als Aufteilungskriterium nicht durch andere ausgetauscht werden können [Let12]. In Abb. 2.8 ist ein Entscheidungsbaum der Tiefe 3 dargestellt.

Ein großer Vorteil von Entscheidungsbäumen ist die einfache und intuitive grafische Darstellung. Die Struktur eines Entscheidungsbaums zeigt bspw. die Wichtigkeit eines Merkmals basierend auf der Tiefe des Merkmals im Baum. Ein Merkmal ist wichtiger, je höher es im Baum eingeordnet ist. Wie Freitas [Fre14] beschreibt, ist ein anderes Maß für die Wichtigkeit des Merkmals, das aus den Entscheidungsbäumen hervorgeht, die Anzahl der Instanzen, die ein Merkmal zur Klassifizierung verwendet.

Entscheidungsbäume können in unterschiedliche Arten eingeteilt werden. Binäre Entscheidungsbäume enthalten genau zwei Verzweigungen (wahr oder falsch). Multivariate Entscheidungsbäume können mehrere Merkmale als Aufteilungskriterium verwenden. Zunächst wird eine Linearkombination dieser Merkmale gelernt und im Anschluss an jedem Knoten als erweitertes Aufteilungskriterium ausgewertet.

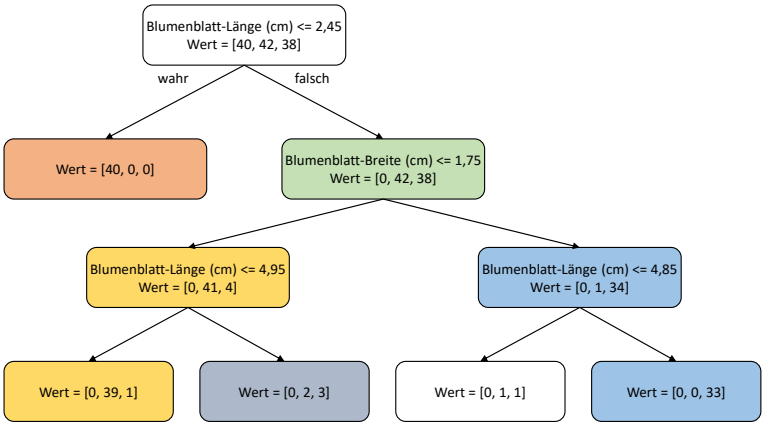


Abbildung 2.8: Entscheidungsbaum der Tiefe 3 am Beispieldatensatz IRIS [Quelle: In Anlehnung an [Bur21b], S. 266].

Entscheidungstabellen: Entscheidungstabellen folgen einer tabellarischen Darstellung, die aus Spalten besteht, die wiederum Regeln beinhalten. Die letzte Zeile der Spalte enthält die Klassifizierung. Es gibt Entscheidungstabellen, die aus sich gegenseitig ausschließenden Regeln bestehen. Ein anderer Typ sind Entscheidungstabellen mit mehreren Treffern. Hier kann der Fall eintreten, dass eine einzelne Instanz durch mehrere Regeln klassifiziert wird.

2.5.3 Untersuchung der Merkmalseigenschaften von Modellen

Erklärbarkeit kann durch die Angabe der Merkmalsrelevanz (engl. feature importance) oder der Merkmalsbeziehung (engl. feature relationship) zum Ergebnis erzeugt werden. Merkmale können für eine Vorhersage auf lokaler und

globaler Ebene relevant sein. Methoden der Merkmalsrelevanz können erweitert werden, um die Beziehung des Merkmals zum Ergebnis des Modells anzuzeigen. Im Folgenden werden Ansätze beschrieben, die Blackbox-Modelle durch die Angabe der Merkmalsrelevanz oder der Merkmalsbeziehung zum Ergebnis von Whitebox- oder Blackbox-Modellen extrahieren.

Merkmalsbeziehung zum Ergebnis: Die Merkmalsbeziehung zum Ergebnis gibt den Einfluss der Merkmale auf das vorhergesagte Ergebnis an. Dazu kann die Art und Weise dargestellt werden, wie sich ein Merkmal auf ein bestimmtes Ergebnis auswirkt. Dabei wird die Abbildung des Effekts der ausgewählten Merkmale aufgedeckt, die entweder linearer oder nicht linearer Natur zwischen dem vorhergesagten Ergebnis und den einzelnen Merkmalen sein kann [Mol20].

Der partielle Abhängigkeitsplot (engl. Partial Dependence Plot (PDP)) zeigt den marginalen Effekt, den ein (s. Abb. 2.10) oder zwei Merkmale (s. Abb. 2.9) auf das vorhergesagte Ergebnis eines Modells haben. Auf diese Weise kann dargestellt werden, ob die Beziehung zwischen dem Ziel und einem Merkmal linear, monoton oder nicht linear ist. Auf ein lineares Regressionsmodell angewendet, zeigen diese stets eine lineare Beziehung [Mol20].

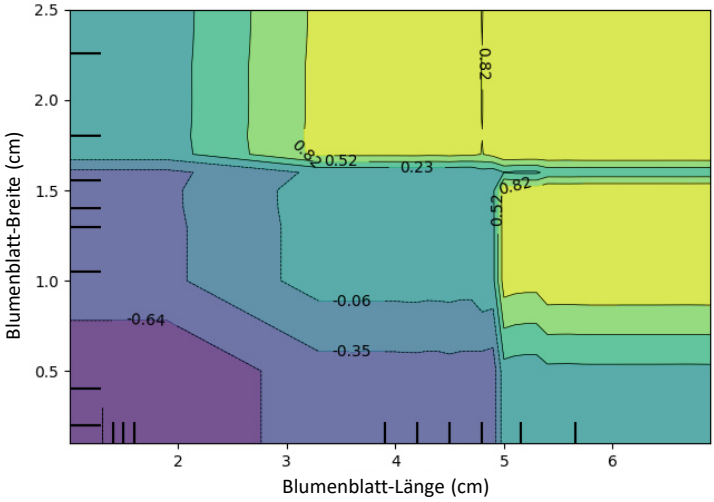


Abbildung 2.9: PDP für die Merkmale *Blumenblatt-Breite* und *Blumenblatt-Länge* der Klasse *Virginica* [Quelle: In Anlehnung an [Bur21b], S. 283].

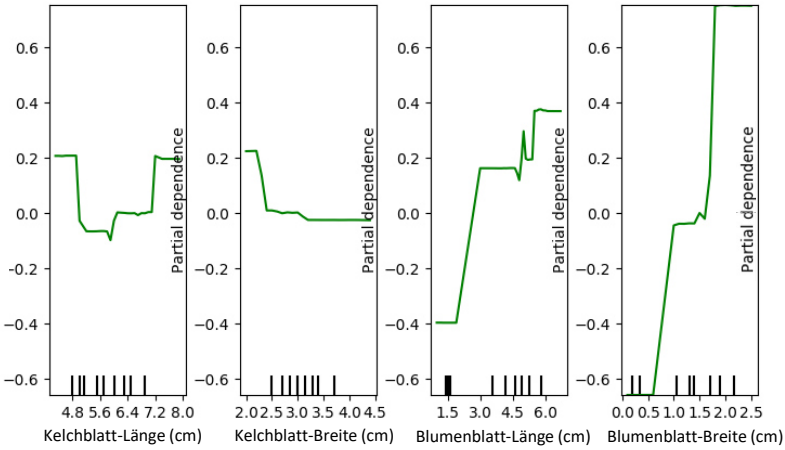


Abbildung 2.10: PDP für die Klasse *Virginica* [Quelle: In Anlehnung an [Bur21b], S. 283].

Merkmalsrelevanz: Eine der grundlegendsten Fragen, die beim Modellverständnis unterstützen kann, zielt darauf ab, herauszufinden, welche Merkmale den größten Einfluss auf die Vorhersage haben. Dieses Konzept wird als *Merkmalsrelevanz* bezeichnet und basiert auf der Idee, dass wichtige Merkmale einen größeren Einfluss auf die Vorhersagen des Modells haben. Erklärbarkeit kann durch die Darstellung der für die Vorhersage am ausschlaggebendsten Merkmale erzeugt werden. Die Merkmalsrelevanz kann für Vorhersagen auf lokaler als auch auf globaler Ebene relevant sein. Wie viel Einfluss ein Merkmal auf die Vorhersage eines Modells hat, kann durch unterschiedliche Arten extrahiert werden. In der Literatur gibt es viele Verfahren der Erklärbarkeit die Merkmalsrelevanzen erzeugen.

Die am häufigsten verwendeten Ansätze zur Extraktion der Merkmalsrelevanz verfolgen das Ziel, einen Merkmalswert zu generieren, der den Effekt

des Merkmals auf die gesamte Vorhersagequalität des Modells darstellt. Darunter fällt auch das zuvor beschriebene Verfahren LIME. Im Folgenden werden unterschiedliche Verfahren der Ansätze beschrieben, die teilweise in den Nutzerstudien in Kapitel 6 verwendet werden.

Das Leave-One-Covariate-Out-Verfahren lernt den Beitrag eines Merkmals, indem es bei der Generierung der Vorhersage weggelassen wird. Dabei wird jedem Merkmal ein lokaler Beitrag zugewiesen, der aggregiert und gemittelt werden kann, um jedem Merkmal auch einen globalen Beitragswert zuzuweisen. Der Ansatz funktioniert gut bei niedrigdimensionalen Daten oder wenn zuvor ein Verfahren der Merkmalsauswahl angewendet wurde [Štr10]. Ein weiteres Verfahren, das einen spieltheoretischen Ansatz verfolgt, nennt sich *SHAP*. Bei der Berechnung wird jeder Wert eines Merkmals entsprechend einer bestimmten Vorhersage gewichtet. Bei der Berechnung werden alle möglichen Kombinationen berücksichtigt. Durch dieses Vorgehen wird die Relevanz eines einzelnen Merkmals berechnet. Ein weiteres Verfahren der Merkmalsrelevanz mit dem Namen Model Class Reliance (MCR) wurde von Smith et al. [Smi20] eingeführt. Dabei wird eine obere und eine untere Grenze festgelegt, in dem sich das Modell für eine Klasse auf ein Merkmal oder eine Gruppe von Merkmalen bezieht. MCR beschreibt also die Abhängigkeit eines Merkmals oder einer Gruppe von Merkmalen zu einer Klasse.

Ein weiterer Ansatz ist die Merkmalsrelevanz basierend auf der *Permutation*. Dabei wird untersucht, inwieweit sich der Vorhersagefehler verändert, wenn ein Merkmal nicht verwendet wird. Zur Messung des Vorhersagefehlers kann eine beliebige Metrik wie z. B. der F1-Score verwendet werden. Um zu vermeiden, dass Merkmale entfernt werden und der Schätzer für jedes Merkmal neu trainiert werden muss, sortiert der Algorithmus die Merkmalswerte zufällig und fügt dem Merkmal ein Rauschen hinzu. Der Vorhersagefehler des neuen Datensatzes wird mit dem Vorhersagefehler des unveränderten Datensatzes verglichen. Fokussiert sich das Modell bei der Vorhersage der Zielvariablen stark auf das geänderte Merkmal, führt dies zu weniger genauen Vorhersagen. Konzentriert sich das Modell bei den Vorhersagen schwach auf das Merkmal, bleibt der Vorhersagefehler unverändert.

2.6 Zusammenfassung

Lernparadigmen des maschinellen Lernens unterteilen sich in drei große Themengebiete: dem überwachten Lernen, dem unüberwachten Lernen und dem semi-überwachten Lernen. Das tiefgehende Lernen ist Teil aller drei Themengebiete. Das Lernparadigma des überwachten maschinellen Lernens, insbesondere der Klassifikation, bildet die Grundlage dieser Arbeit. Die Verfahren im Bereich der Klassifikation werden entsprechend ihrer Ausgabe in zwei Kategorien unterteilt. Zum einen können die Verfahren Blackbox-Modelle erzeugen, die nicht direkt nachvollziehbar sind. Zum anderen können die Verfahren interpretierbare Modelle (Whitebox-Modelle) erzeugen, die dadurch auch zur Erzeugung von Erklärungen verwendet werden können. Blackbox-Modelle eignen sich ebenfalls zur Erzeugung von Erklärungen, jedoch benötigen diese dazu spezielle Erweiterungen, wie z. B. in Form eines Surrogates.

Durch die hohe Genauigkeit der Blackbox-Modelle steigt zugleich deren Komplexität. Die erhöhte Komplexität erfordert neue Konzepte, die die Modelle für Anwender nachvollziehbar gestalten, um Blackbox-Modelle zu beleuchten. Dieses Forschungsgebiet wird als *erklärbare Künstliche Intelligenz* bzw. *erklärbares maschinellen Lernens* bezeichnet. Das Gebiet ist kein neues Forschungsfeld, hat jedoch in den letzten Jahren stark an Interesse – sowohl in der Forschung als auch in der Industrie – gewonnen. Durch den Einsatz erklärbarer KI können viele unterschiedliche Domänen wie die Medizin, das Finanz- und Justizwesen sowie der Automobilsektor profitieren. Verfahren der Erklärbarkeit lassen sich grob in modell-agnostische, modell-spezifische, lokale bzw. globale und Post- bzw. Ante-hoc-Verfahren einteilen. Die Ergebnisse der Verfahren sind Erklärungen.

Einzelne Vorhersagen und Modelle können bspw. von Domänenexperten detailliert verstanden und von Entwicklern gezielt auf Fehler untersucht werden. Eventuelle Verzerrungen des Modells werden frühzeitig erkannt. Allgemeine Vorbehalte gegenüber Systemen, die auf den Verfahren basieren, werden geringer. Zukünftig wird es eine rechtliche Forderung sein, als betroffene Person eine Erklärung für die Entscheidung eines Modells zu erhalten.

Welche Anforderungen eine Erklärung erfüllen muss, ist noch nicht eindeutig definiert. Bislang fehlen einheitliche Definitionen der Erklärbarkeit. In der vorliegenden Arbeit wird Erklärbarkeit als die Fähigkeit definiert, das Modell oder dessen Ergebnis nachvollziehen zu können. Der Anwender muss nicht zwingend ein Domänenexperte oder Entwickler sein. Die zu lösende Aufgabe muss vom Anwender verstanden werden und somit auch die Erklärung.

Wie in diesem Kapitel aufgezeigt wurde, gibt es unterschiedliche Arten von Erklärungen: lokale oder globale faktische, lokale kontrafaktische oder prototypische lokale oder globale Erklärungen. Kontrafaktische Erklärungen beschreiben, wie eine getroffene Entscheidung anhand der gegebenen Merkmale verändert werden kann, um möglicherweise eine bevorzugte Entscheidung zu erhalten. Oftmals kann auch ohne die Implementierung von Verfahren, die Erklärungen erzeugen, Nachvollziehbarkeit geschaffen werden. Ein Aspekt, der bislang zu wenig Beachtung gefunden hat, ist die Qualität der Daten, mit denen die Modelle trainiert werden. Sofern die Daten verrauscht, durch eine nicht repräsentative Erhebung verzerrt oder falsch sind, werden ebenso die darauf resultierenden Modelle nicht korrekte Vorhersagen treffen können (auch bei einer vermutlich hohen Genauigkeit der Modelle). Ein wichtiges Werkzeug zur Erzeugung von Erklärbarkeit innerhalb der Modelle ist die Regularisierung. Diese fügt der Zielfunktion eines Modells einen Strafterm hinzu, um die Komplexität des Modells zu reduzieren. Diese Technik wird für das in Kapitel 4 entworfene Verfahren verwendet.

Mittlerweile gibt es eine Vielzahl von Verfahren zur Erzeugung von Erklärungen. Das Kapitel Basis-Verfahren beschreibt einige grundlegende Verfahren zur Erzeugung von Erklärungen. Diese sind ein Baustein für die in Kapitel 4 und Kapitel 5 entworfenen Verfahren sowie auch teilweise in Kapitel 6 den Nutzerstudien. Inwieweit die Erklärungen jedoch tatsächlich dem Anwender bei der Nachvollziehbarkeit helfen, wird in der Regel nicht weiter betrachtet. Neben den Verfahren kommt Evaluierungen wie bspw. durch Nutzerstudien eine besondere Bedeutung zu. Mit diesen kann untersucht werden, inwieweit

die erzeugten Erklärungen bei der Nachvollziehbarkeit der Ergebnisse unterstützen. Zur gezielten Optimierung von Erklärungen werden Ontologien eingesetzt. Diese werden dazu verwendet, um Wissen einfach darstellen zu können. Dieser Aspekt wurde in Bezug auf Erklärungen in Nutzerstudien bislang nur wenig untersucht.

Nachdem in diesem Kapitel ein Überblick über die grundlegende Methodik maschineller Lernverfahren und der Erklärbarkeit dieser gegeben wurde, wird im nachfolgenden Kapitel 3 ein Vorgehensmodell zur Extraktion verschiedener Erklärarten gegeben. Dies ist für die weitere Einordnung der Verfahren, die in Kapitel 4 und Kapitel 5 entworfen werden essenziell.

3 Vorgehensmodell zur Extraktion unterschiedlicher Arten von Erklärungen

In diesem Kapitel wird das theoretische Vorgehensmodell beschrieben, das unterschiedliche Arten zur Extraktion von Erklärungen definiert und das Rahmenwerk der vorliegenden Arbeit bildet. Die Arbeiten basieren auf der Veröffentlichung von Burkart et al. [Bur21b].

Zunächst wird erläutert, welche Arten von Erklärungen im Bereich überwachter maschineller Lernverfahren für den Teilbereich der Klassifikation erzeugt werden können. Insgesamt wird dabei zwischen fünf verschiedenen Arten von Erklärungen unterschieden. Die Problemdefinition der Klassifikation in Abschnitt 2.1.2.1 bildet die Ausgangsbasis. Darauf aufbauend werden die fünf Arten der Erklärungen eingeführt. Die Verfahren zur Erzeugung von Erklärungen aus der Literatur werden direkt in die jeweiligen Abschnitte der Erklärungen eingeordnet. In Kapitel 4 und Kapitel 5 werden Verfahren entworfen, die ebenfalls in das Vorgehensmodell eingeordnet werden.

3.1 Das Vorgehensmodell

In Abbildung 3.1 ist das Vorgehensmodell zur Extraktion fünf unterschiedlicher Arten von Erklärungen dargestellt. Der klassische Pfad des überwachten Lernens ohne die Extraktion einer Erklärung ist unter (a) dargestellt. Die Teile (b)-(f) zeigen die fünf Arten an: die Teile (b)-(d) stellen Modell- bzw. globale Erklärungen dar und die Teile (e) und (f) Instanz- bzw. lokale Erklärungen.

Jede der fünf Arten von Erklärungen enthält mindestens die Bausteine Lernalgorithmus, Modell und Erklärung. Der Lernalgorithmus erstellt zusammen mit den Trainingsdaten $\mathcal{D}$ das Modell. Erst wenn das Modell vom Lernalgorithmus erstellt wurde, kann daraus eine entsprechende Erklärung extrahiert werden. Der Vorgang (b) in Abbildung 3.1 zeigt die Extraktion direkter globaler bzw. von Modell-Erklärungen und wird in Abschnitt 3.6.1 beschrieben. Der Vorgang (c) demonstriert die Generierung interpretierbarer Modelle (s. Abschnitt 3.4) und (d) die Generierung von Blackbox-Erklärungen mithilfe von Surrogaten (s. Abschnitt 3.5.1). Die Erklärung kann entweder direkt wie bei (e) (s. Abschnitt 3.6.2) oder mithilfe eines lokalen Surrogates wie bei (f) (s. Abschnitt 3.5.2) extrahiert werden.

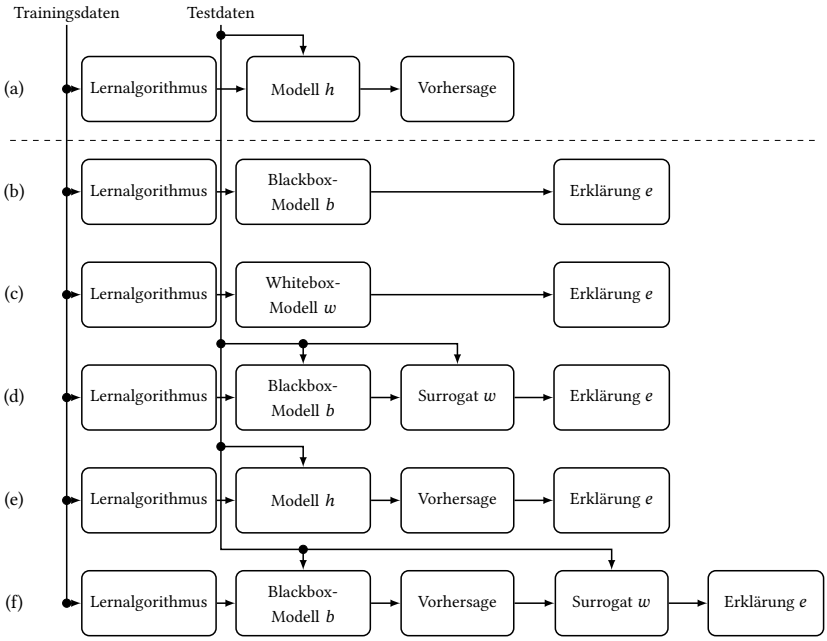


Abbildung 3.1: Vorgehensmodell zur Extraktion unterschiedlicher Erklärungen [Quelle: In Anlehnung an [Bur21b], S. 252].

Die nachfolgenden Abschnitte 3.2-3.6 erläutern die einzelnen Arten zur Extraktion einer Erklärung (Vorgänge (b)-(f) in Abb. 3.1.

3.2 Die Problemdefinition des überwachten maschinellen Lernens als Ausgangsbasis

In Abschnitt 2.1.2.1 wurde das Problem 1 des überwachten maschinellen Lernens beschrieben. Die Lösung des Problems führt in der Regel zu Blackbox-Modellen, die ohne die Hinzunahme einer weiteren Lösung für den Anwender nicht nachvollziehbar bleiben. Diese werden durch die Modifikation oder die Erweiterung von Gleichung (2.1) erreicht. Im Allgemeinen können diese Probleme in *Modell-* (vgl. Abb. 3.1(b)–(d)) und *Instanz-Erklärungen* (vgl. Abb. 3.1(e)–(f)) eingeteilt werden. Modell-Erklärungen generieren Erklärungen basierend auf dem trainierten Modell und zielen darauf ab, Einblicke in das Innenleben des gesamten Modells geben zu können. Im Gegensatz dazu versuchen Instanz-Erklärungen, eine Vorhersage y für eine einzelne Dateninstanz x zu generieren. Die erzeugten Erklärungen sind somit nur für x und dessen nähere Umgebung gültig.

In den Abschnitten 3.2 - 3.6 werden die verschiedenen Wege zur Extraktion von Erklärungen formal definiert. Innerhalb der einzelnen Abschnitte wird der Stand von Forschung und Technik zu Verfahren aus der Literatur direkt eingeordnet. Die Tabellen zur Einordnung der Verfahren in den einzelnen Abschnitten sind alle gleich aufgebaut. Diese enthalten neben dem Verfahrensnamen die Lernaufgabe, die entweder nur die Klassifikation oder zusätzlich auch die Regression adressieren kann. Zudem wird das Ergebnismodell angegeben, das darüber Aufschluss gibt, welchem Modell die Erklärung entstammt. Sämtliche Tabellen, ausgenommen in Abschnitt 3.4, verfügen über eine weitere Tabellenspalte, dem *Umfang*. Dieser gibt darüber Aufschluss, ob die Verfahren entweder modell-agnostisch oder modell-spezifisch sind. Interpretierbare Modelle sind per se modell-spezifisch und enthalten diese Angabe daher nicht.

3.3 Das Resultat - Die Erklärung

Die Lernalgorithmen trainieren mit den Trainingsdaten das Modell, und das Modell generiert die Vorhersagen für neue Dateninstanzen. Eine Erklärfunktion $e \in \mathcal{E}$ versucht, für Anwender nachvollziehbare Erklärungen zu generieren, die entweder das Modell und/oder die Vorhersage erklären. Für jede Art von Erklärung ist eine Funktion zur Generierung der eigentlichen Erklärung notwendig, um das verwendete Modell oder die erhaltene Vorhersage nachzuvollziehen. Der Aufbau einer Erklärung wurde bereits in Abschnitt 2.3 vorgestellt. Die Erklärfunktion stellt zusätzliche Mittel zur besseren Nachvollziehbarkeit des Modells bereit, wie bspw. die Angabe der Merkmalsrelevanz oder bestimmte Visualisierungen. Erklärungen stützen sich immer auf ein gelerntes Modell, das entweder eine Whitebox oder eine Blackbox sein kann. Demnach benötigen auch interpretierbare Modelle eine Erklärfunktion, um Erklärungen wie z. B. die Visualisierung eines Entscheidungsbaumes zu generieren. Das Vorgehen zur Generierung einer Erklärung wird in Gleichung (3.1) gegeben.

Problem 2 (Generierung der Erklärung)**.** Eine Erklärfunktion e ist definiert als

$$e : (\mathcal{X} \to \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathcal{E} \tag{3.1}$$

die zum einen das Vorhersagemodell (entweder Blackbox oder interpretierbar) und zum anderen einen Datensatz als Eingabe verwendet, um eine Erklärung zu generieren. Die Erklärung e gehört zu der Menge aller Erklärungen $\mathcal{E}$. Dabei werden zwei Arten der Generierung von Erklärungen unterschieden:

Global: Die Extraktion einer globalen Erklärung erfolgt basierend auf einem Modell, die repräsentativ für einen Datensatz ist $\mathcal{D}$, $e(b, \mathcal{D})$ bspw. bei einem Blackbox-Modell (siehe Abb. 3.1(b)) oder $e(w, D)$ für ein interpretierbares Modell (siehe Abb. 3.1 (c) und (d)).

Lokal: Die Extraktion einer lokalen Erklärung erfolgt für eine einzelne Instanz x und die dazugehörige Vorhersage y, bspw. $e(b, (x, y))$ oder $e(w, (x, y))$.

3.4 Interpretierbare Modelle

Zunächst wird das Problem der interpretierbaren Modelle betrachtet, das in Abb. 3.2 separat dargestellt ist.

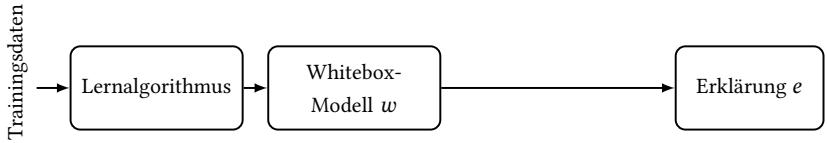


Abbildung 3.2: Trainieren von interpretierbaren Modellen [Quelle: In Anlehnung an [Bur21b], S. 264].

Durch das Lösen des Problems 3 wird versucht, ein interpretierbares Modell aus dem Hypothesenraum interpretierbarer Modelle $\mathcal{W}$ zu lernen. Dazu trainiert der Lernalgorithmus anhand der Trainingsdaten ein interpretierbares Modell $w \in \mathcal{W}$. Im Anschluss daran wird mit der Erklärfunktion aus Gleichung (3.1) die Erklärung erzeugt. Dieser Weg wird auch als *Ante-hoc-Erklärbarkeit* bezeichnet werden (s. Abschnitt 2.2.6). Typische Beispiele sind das Lernen kleiner Entscheidungsbäume, -listen oder die Verwendung linearer Modelle. Hier ist zu erwähnen, dass die Erklärungen für bspw. Entscheidungsbäume per se nicht unbedingt immer nachvollziehbar sind da diese zu groß werden können.

Problem 3 (Whitebox-Modell (Interpretierbares Modell)). Die Anpassung des Problems 1 führt zu dem Optimierungsproblem

$$w^* = \arg\min_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} S\big(w(x_i), y_i\big)\,, \tag{3.2}$$

mit $(x_i, y_i) \in \mathcal{D}$.

Bei der Einteilung der interpretierbaren Modelle wird zwischen zwei Arten unterschieden: zum einen zwischen *inherent interpretierbaren Modellen* und zum anderen zwischen explizit für die Erklärbarkeit *optimierten interpretierbaren Modellen*. Beide Arten werden nachfolgend beschrieben da diese die Grundlage der Surrogaten-Modelle bilden.

3.4.1 Inherente Whitebox-Modelle

Eine genauere Betrachtung einiger klassischer Algorithmen des überwachten maschinellen Lernens zeigt, dass diese bereits interpretierbare Modelle generieren, ohne dass diese explizit dafür optimiert wurden. Das Training der Lernalgorithmen wurde dementsprechend nicht darauf optimiert, nachvollziehbar zu sein. Klassischerweise sind die Algorithmen auf eine hohe Genauigkeit optimiert, die Erklärbarkeit ist dabei ein Nebenaspekt. Diese ist dabei natürlich gegeben und daher inherent. Die resultierenden Modelle können direkt bspw. durch Visualisierungen dargestellt werden, um Ergebnisse des Modells nachvollziehen zu können. Ein wichtiger Aspekt bei dieser Art von Modellen ist, dass diese möglichst klein und einfach für Anwender zu verstehen sind. Ein Beispiel für ein inherent interpretierbares Modell ist ein Entscheidungsbaum der Tiefe 2 oder 3. Dieser Kategorie lassen sich weitere Verfahren aus der Literatur zuordnen. Die Basisverfahren wurden bereits in Abschnitt 2.1.2.1 beschrieben. Ein Überblick der Verfahren aus der Literatur wird in Tabelle 3.1 gegeben.

Tabelle 3.1: Verfahren zur Erzeugung inherent interpretierbarer Modelle [Quelle: In Anlehnung an [Bur21b], S. 269].

Verfahren	Lernaufgabe	Ergebnismodell	Referenz
GAMs	Klassifikation	Lineares Modell	Lou et al. [Lou12]
CN2	Klassifikation	Regelbasiert	Clark et al. [Cla89]
RIPPER	Klassifikation	Regelbasiert	Cohen [Coh95]
Re-RX	Klassifikation	Regelbasiert	Setiono et al. [Set08]
C5.0R	Klassifikation	Regelbasiert	Ustun et al. [Ust16]
AntMiner+	Klassifikation	Regelbasiert	Martens et al. [Mar07]
cAntMinerPB	Klassifikation	Regelbasiert	Otero et al. [Ote16]
CART	Klassifikation & Regression	Baumbasiert	Breiman [Bre17]
ID2of3	Klassifikation & Regression	Baumbasiert	Craven et al. [Cra96]
ID3	Klassifikation & Regression	Baumbasiert	Quinlan [Qui86]
C4.5	Klassifikation & Regression	Baumbasiert	Quinlan [Qui14]
C5.0T	Klassifikation & Regression	Baumbasiert	Ustun et al. [Ust16]
Bayes'sches Netz	Klassifikation	Bayes'sches Netz	Friedman et al. [Fri97]
Lineare Regression (Lasso)	Regression	Lineares Modell	Tibshirani [Tib96]
Lineare Regression (LARS)	Regression	Lineares Modell	Efron et al. [Efr04]
Logistische Regression	Klassifikation	Lineares Modell	Berkson [Ber53]
AOT	Klassifikation	Baumbasiert	Si et al. [Si13]

3.4.2 Optimierte Whitebox-Modelle

Der Vorgang der optimierten Whitebox-Modelle versucht, den Aspekt der Erklärbarkeit in das Design des Trainings miteinzubeziehen. Im Gegensatz zu den *inherent interpretierbaren Modellen* erlauben die *optimierten interpretierbaren Modelle*, den Grad der Erklärbarkeit zu kontrollieren. Das bedeutet, dass die resultierenden Modelle gezielt dazu verwendet werden können, eine erhöhte Erklärbarkeit für den Anwender zu generieren. Dies kann die Genauigkeit der Modelle beeinträchtigen und wird daher oft als Trade-off zwischen Genauigkeit und Erklärbarkeit bezeichnet. Die Verfahren lassen sich grob in die Kategorien Entscheidungsbäume, Entscheidungsregeln und -sätze sowie lineare Modelle einteilen. Die Grundkonzepte dieser Verfahren wurden bereits in Abschnitt 2.1.2.1 beschrieben.

Um die Zuordnung von Verfahren aus der Literatur aufzuzeigen, wird nachfolgend in Tabelle 3.2 ein Überblick über spezielle Verfahren gegeben.

Tabelle 3.2: Optimierte Verfahren zur Erzeugung interpretierbarer Modelle [Quelle: In Anlehnung an [Bur21b], S. 270.]

Verfahren	Lernaufgabe	Ergebnis	Referenz
SLIM	Klassifikation	Regelbasiert	Ustun et al. [Ust14]
TILM	Klassifikation	Regelbasiert	Ustun et al. [Ust16]
PILM	Klassifikation	Lineares Modell	Ustun et al. [Ust16]
RiskSLIM	Klassifikation	Regelbasiert	Ustun et al. [Ust17]
Boolean Rules	Klassifikation	Regelbasiert	Su et al. [Su15]
BOA	Klassifikation	Regelbasiert	Wang et al. [Wan15c]
ORC	Klassifikation	Regelbasiert	Bertsimas et al. [Ber11]
BLM	Klassifikation	Regelbasiert	Letham et al. [Let12]
BRL	Klassifikation	Regelbasiert	Letham et al. [Let15]
(S)BRL	Klassifikation	Regelbasiert	Yang et al. [Yan17]
FRL	Klassifikation	Regelbasiert	Wang et al. [Wan15a]
IDS	Klassifikation	Regelbasiert	Lakkaraju et al. [Lak17]
BRS	Klassifikation	Regelbasiert	Wang et al. [Wan16]
OT-SpAM	Klassifikation	Baumbasiert	Wang et al. [Wan15b]
Baum-Reg.	Klassifikation	Baumbasiert	Wu et al. [Wu18]
Ortho. Reg.	Klassifikation	Baumbasiert	Schaaf et al. [Sch19a]
NDT	Klassifikation	Baumbasiert	Balestriero [Bal17]
DNDT	Klassifikation	Baumbasiert	Yang et al. [Yan18b]
LP relaxation	Klassifikation	Regelbasiert	Malioutov et al. [Mal17]
1R	Klassifikation	Regelbasiert	Holte [Hol93]
TLBR	Klassifikation	Regelbasiert	Su et al. [Su16]
CPAR	Klassifikation	Regelbasiert	Yin et al. [Yin03]

3.5 Surrogate Modellanpassung

Während die Verwendung von interpretierbaren Modellen für viele Probleme geeignet sein kann, können sich diese laut Ribeiro et al. [Rib16a] nachteilig auf die Flexibilität, die Genauigkeit und die Wiederverwendbarkeit des Modells auswirken, da diese Verfahren modell-spezifisch sind. An dieser Stelle wird

allerdings auch angemerkt, dass es viele Befürworter gibt, die sich klar für die Verwendung von interpretierbaren Modellen im Gegensatz zu Blackbox-Modellen mit Erklärungen aussprechen [Rud18]. Rudin [Rud18] kritisiert den Anstieg von Surrogat-Modelllösungen für folgenschwere Entscheidungen wie im medizinischen oder juristischen Bereich.

Um die sogenannte *Post-hoc-Erklärbarkeit* zu generieren, wird das Blackbox-Modell weiterhin für die Vorhersagen verwendet, da dieses in der Regel eine hohe Genauigkeit hat. Zusätzlich zu dem Blackbox-Modell wird ein interpretierbares Stellvertretermodell, nachfolgend nur noch Surrogat oder Surrogat-Modell genannt, erzeugt. Dadurch wird die Nachvollziehbarkeit des Blackbox-Modells erzeugt. Das Surrogat-Modell übersetzt das Vorhersagemodell in ein approximiertes Modell [Hen14].

Im nächsten Schritt wird zunächst die Problemdefinition der surrogaten Modellanpassung gegeben.

Problem 4 (Surrogate Modellanpassung)**.** Die surrogate Modellanpassung generiert aus dem Blackbox-Modell ein approximiertes interpretierbares Modell, indem versucht wird, die folgende Gleichung zu lösen:

$$w^* = \arg\min_{w \in \mathcal{W}} \frac{1}{|\mathcal{X}|} \sum\nolimits_{x \in \mathcal{X}} S\big(w(x), b(x)\big)\,. \tag{3.3}$$

Dabei wird zwischen zwei Arten von Surrogat-Modellen unterschieden:

Global: Das Surrogat-Modell w approximiert das Blackbox-Modell b auf der gesamten Trainingsmenge, bspw. ist $X = \{x_1, \ldots, x_n\}$ Teil des Trainingsdatensatzes $\mathcal{D}$. Alternativ kann $\mathcal{X}$ ein Beispieldatensatz sein, der die Verteilung der Eingabedaten des Modells b repräsentiert.

Lokal: Das Surrogat-Modell w approximiert das Blackbox-Modell um die Instanz x, das als $\mathcal{X} = \{x' | x' \in N(x)\}$ definiert ist, wobei $\mathcal{N}$ in der Nachbarschaft von x liegt.

In Gleichung (3.3) ist das Maß S ein Treuemaß, das darüber Aufschluss gibt, wie gut das Surrogat-Modell w das Blackbox-Modell b in den Entscheidungen nachahmt.

Die Vorgänge zur Generierung der Erklärung basierend auf den globalen und lokalen Surrogaten sind in Abb. 3.1 (d) und (f) abgebildet. Das globale Surrogat versucht, alle Vorhersagen des Blackbox-Modells mit einer möglichst hohen Genauigkeit nachzuahmen. Ein vereinfachtes Beispiel hierfür ist, einen Entscheidungsbaum auf dem Trainingsdatensatz $\mathcal{D}' = \{(x_1, b(x_1)), \ldots, (x_n, b(x_n))\}$ zu trainieren. Das Surrogat ist auf den Vorhersagen $b(x_i)$ des Blackbox-Modells trainiert, x_i sind die Merkmalsvektoren des Trainingsdatensatzes. Ein lokales Surrogat ist nur in der Nähe der betrachteten Instanz x gültig. Dabei kann folglich nur ein lokales Verständnis über die Blackbox erlangt werden.

Die Verfahren aus der Literatur zur Erzeugung globaler als auch lokaler Surrogate lassen sich grob in die Kategorien Entscheidungsbäume, Entscheidungsregeln und lineare Modelle einteilen. Ein Überblick über die Verfahren aus der Literatur wird in Tabelle 3.3 und Tabelle 3.4 gegeben.

3.5.1 Globale Surrogate

In diesem Abschnitt werden globale Surrogate beschrieben, die nachvollziehbare Modelle lernen, um die Vorhersagen eines Blackbox-Modells nachzuahmen. Die Übersicht über die verschiedenen Ansätze kann Tabelle 3.3 entnommen werden. Der Prozess der Anpassung eines globalen Surrogates ist in Abb. 3.3 dargestellt. Zusammen mit den Trainingsdaten und dem Lernalgorithmus wird das Blackbox-Modell erstellt. Das Blackbox-Modell kann sodann zusammen mit den Testdaten zur Generierung der Vorhersagen verwendet werden. Das Surrogat-Modell kann entweder parallel neben dem Training der Blackbox oder nach dem Training abgeleitet werden. Das Surrogat wird basierend auf den Entscheidungen des Blackbox-Modells und den Traningsdaten angepasst. Das Surrogat erhält zur Erzeugung von Erklärungen auch die Testdaten. Im Anschluss daran wird mit der Erklärfunktion aus Gleichung (3.1) die Erklärung erzeugt.

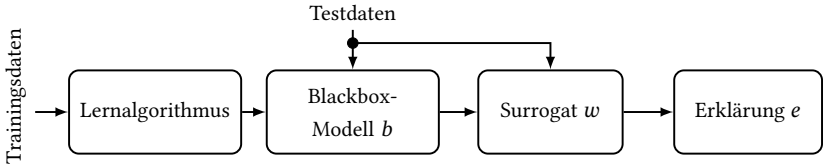


Abbildung 3.3: Anpassung eines globales Surrogates entsprechend Abb. 3.1 (d) [Quelle: In Anlehnung an [Bur21b], S. 270].

Tabelle 3.3: Verfahren aus der Literatur zur Erzeugung globaler Surrogat-Modelle [Quelle: In Anlehnung an [Bur21b], S. 271].

Verfahren	Umfang	Ergebnismodell	Referenz
sp-Lime	Agnostisch	Lineares Modell	[Rib16b]
k-Lime	Agnostisch	Lineares Modell	[Hal17a]
Tree Merging	Spezifisch	Baumbasiert	[And13]
Decision Tree Extract	Spezifisch	Baumbasiert	[Bas17]
Soft Decision Tree	Agnostisch	Baumbasiert	[Hin17]
Binary Decision Tree	Agnostisch	Baumbasiert	[Yan18a]
Probabilistic interpretation	Spezifisch	Baumbasiert	[Sch07]
EM min. Kullback	Spezifisch	Baumbasiert	[Har16]

3.5.2 Lokale Surrogate

In diesem Abschnitt werden Verfahren betrachtet, die ein lokales Surrogat-Modell generieren. Im Gegensatz zu den in Abschnitt 3.5.1 beschriebenen globalen Surrogaten ist das lokale Surrogat nur für eine bestimmte Instanz und deren Umgebung gültig. LIME (s. Abschnitt 2.5) [Rib16a] ist beispielsweise ein Verfahren, dass der Klasse der lokalen Surrogate zugeordnet ist. Zur Einordnung der Verfahren aus der Literatur, wird eine Übersicht über Verfahren im Bereich der lokalen Surrogate in Tabelle 3.4 gegeben.

Tabelle 3.4: Verfahren zur Erzeugung lokaler Surrogat-Modelle [Quelle: In Anlehnung an [Bur21b]), S.271].

Verfahren	Lernaufgabe	Ergebnis	Referenz
LIME	Klassifikation / Regression	Lineares Modell	[Rib16a]
aLime	Klassifikation	Lineares Modell	[Rib18]
MES	Klassifikation	Lineares Modell	[Tur16]
MUSE	Klassifikation	Lineares Modell	[Lak19]
LORE	Klassifikation	Lineares Modell	[Gui18b]
MAPLE	Klassifikation	Lineares Modell	[Plu18]
Kernel SHAP	Klassifikation	Lineares Modell	[Lun17]
Linear SHAP	Klassifikation/ Regression.	Lineares Modell	[Lun17]

3.6 Direkte Extraktion von Erklärungen

Nachfolgend werden Verfahren beschrieben, die eine Erklärung entweder direkt lokal oder global generieren können. Der Unterschied zu den Surrogat-Modellen besteht darin, dass die Erklärung ohne ein Zwischenmodell direkt aus dem Blackbox-Modell generiert werden kann. Demnach kann direkt die Erklärfunktion aus Gleichung (3.1) auf das jeweilige Blackbox-Modell angewendet werden. Die direkte Extraktion einer Erklärung aus einem interpretierbaren Modell wäre wiederum der Typ *interpretierbare Modelle* aus Abschnitt 3.4.

3.6.1 Direkte Extraktion einer globalen Erklärung

Im Mittelpunkt dieses Abschnitts stehen Verfahren, die direkt eine Erklärung extrahieren können. Zusammen mit den Trainingsdaten und dem Lernalgorithmus wird das Blackbox-Modell erstellt. Im Gegensatz zu den Surrogat-Modellen wird allerdings kein Zwischenmodell angepasst, um die Erklärfunktion darauf anzuwenden, sondern direkt auf die Blackbox angewendet. Wie in

Abb. 3.4 verdeutlicht, sind die Erklärungen unabhängig von bestimmten Modellvorhersagen, da diese versuchen, bestimmte Eigenschaften des Blackbox-Modells aufzudecken. Die Verfahren zur Erzeugung globaler Erklärungen sind in Tabelle 3.5 aufgelistet.

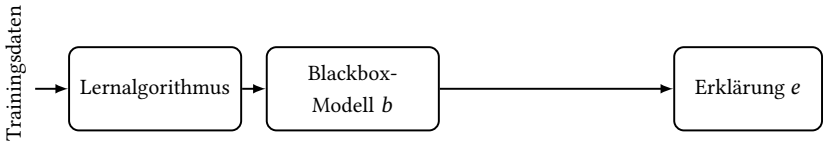


Abbildung 3.4: Generierung einer globaler Erklärung entsprechend Abb. 3.1 (b) [Quelle: In Anlehnung an [Bur21b], S.281].

Tabelle 3.5: Überblick über die direkte Extraktion einer globalen Erklärung [Quelle: In Anlehnung an [Bur21b], S.280].

Verfahren	Lernaufgabe	Umfang	Ergebnismodell	Referenz
RF Merkmalsrelevanz	Klassifikation	Spezifisch	baumbasiert	[Hal17a]
Sparsity Constraints	-	-	-	[Ust14]
Korrelationsgraph	-	Agnostisch	-	[Hal17b]
Residual Analysis	-	Agnostisch	-	[Hal17b]
Autoencoder	-	Agnostisch	-	[Hal17b]
PCA	-	Agnostisch	-	[Hal17b]
MDS	-	Agnostisch	-	[Hal17b]
t-SNE	-	Agnostisch	-	[Hal17b]
Nomograms	Klassifikation	Agnostisch	lineares Modell	[Rob08]
SOM	-	Agnostisch	-	[Mar08]
Quasi Regression	Klassifikation	Agnostisch	-	[Jia02]
EXPLAINER global	Klassifikation	Agnostisch	-	[Sub07]
GSA	Klassifikation	Agnostisch	-	[Cor11]
GOLDEN EYE	Klassifikation	Agnostisch	-	[Hen14]
GFA	Klassifikation	Agnostisch	-	[Adl18]
ASTRID	Klassifikation	Agnostisch	-	[Hen17]
PDP	Klassifikation	Agnostisch	-	[Gol15]
IME	Klassifikation Regression	Agnostisch	-	[Boh17]
Monotonicity Constraints	-	-	-	[Fre14]
Prospector	Klassifikation	Agnostisch	-	[Kra16]
LOCO	Klassifikation	Agnostisch	-	[Str10]

3.6.2 Direkte Extraktion einer lokalen Erklärung

Lokale Erklärungen sind nur in der Nähe einer bestimmten Vorhersage gültig, wie die Abb. 3.5 zeigt. Zunächst wird wieder zusammen mit den Trainingsdaten und dem Lernalgorithmus das Modell trainiert. Das Modell kann dabei entweder eine Blackbox oder ein interpretierbares Modell sein. Zusammen mit den Testdaten erzeugt das Modell h die Vorhersage. Durch die Anwendung der Erklärfunktion wird widerum einer Erkläung e erzeugt. Dieser Abschnitt gibt einen Überblick über die jeweiligen Verfahren des Teilbereichs. Die Verfahren zur Erzeugung lokaler Erklärungen sind in Tabelle 3.6 aufgelistet.

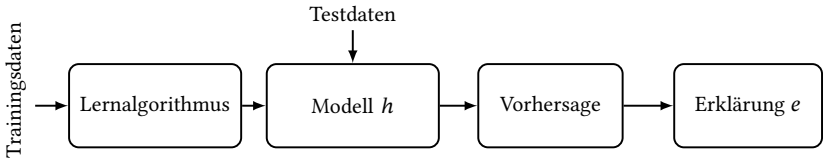


Abbildung 3.5: Generierung einer lokalen Erklärung entsprechend Abb. 3.1 (e) [Quelle: In Anlehnung an [Bur21b], S. 284].

Tabelle 3.6: Verfahren zur Erzeugung direkter lokaler Erklärungen [Quelle: In Anlehnung an [Bur21b], S. 286].

Verfahren	Lernaufgabe	Umfang	Ergebnis	Referenz
ICE Plots	Klassifikation	Agnostisch	-	[Gol15]
EXPLAINER local	Klassifikation	Agnostisch	-	[Sub07]
Border Klassifikation	Klassifikation	Spezifisch	nicht lineares Modell	[Bar09]
Cloaking	Klassifikation	Agnostisch	-	[Che15]
Tweaking Recommendation	Klassifikation	Spezifisch	nicht lineares Modell	[Tol17]
Nearest Neighbor	Klassifikation	Agnostisch	-	[Fre14]
SVM Recommendations	Klassifikation	Spezifisch	nicht lineares Modell	[Bar09]
PVM	Klassifikation	Agnostisch	-	[Bie11]
Bayesian Case Model	Klassifikation	Agnostisch	-	[Kim14]
MMD-critic	Klassifikation	Agnostisch	-	[Kim16]
Influence Functions	Klassifikation	Agnostisch	-	[Koh17]
Local Gradients	Klassifikation	Agnostisch	-	[Bae10]
LOCO	Klassifikation	Agnostisch	-	[Lei18]
QII	Klassifikation	Agnostisch	-	[Dat16]
SENN	Klassifikation	-	Lineares Modell	[Mel18]
VEM	Klassifikation	Spezifisch	nicht lineares Modell	[Hen16]
Treeinterpreter	Klassifikation & Regr.	Spezifisch	Baumbasiert	[Hal17b]
ICM	Klassifikation	Spezifisch	nicht lineares Modell	[Ved17]
DeepLIFT	Klassifikation	Spezifisch	nicht lineares Modell	[Shr16]
SmoothGrad	Klassifiaktion	Spezifisch	nicht lineares Modell	[Smi17]
Interior Gradients	Klassifikation	Spezifisch	nicht lineares Modell	[Sun16]
explainVis	Klassifikation	Spezifisch	Lineares Modell	[Rob08]
Deep Taylor	Klassifikation	Spezifisch	nicht lineares Modell	[Mon17]
LRP	Klassifikation	Spezifisch	nicht lineares Modell	[Sam17]
SA	Klassifikation	Spezifisch	nicht lineares Modell	[Sam17]

3.6.3 Einordnung der Erklärtypen entsprechend der Taxonomie

In Kapitel 2.2.6 wurden die Dimensionen der Taxonomie im Bereich der Erklärbarkeit eingeführt. Die nachstehende Tabelle 3.7 fasst die Dimensionen bezüglich der Problemstellungen im Bereich des überwachten maschinellen Lernens zusammen. Das Häkchen (✓) symbolisiert, dass die bestimmte Dimension von dem Konzept erfüllt wird. Das Kreuz (x) bedeutet, dass die Dimension für das Konzept nicht anwendbar ist, und der Punkt (∘) beschreibt, dass der Ansatz teilweise innerhalb der Dimension anwendbar ist.

Tabelle 3.7: Dimensionen des erklärbaren maschinellen Lernens [Quelle: In Anlehnung an [Bur21b], S. 257].

Ansatz	Whitebox-Modell	Surrogat-Modell	Direkte Erklärungen
ante-hoc	✓	x	x
post-hoc	x	✓	✓
lokal	∘	✓	✓
global	✓	✓	✓
spezifisch	✓	✓	✓
agnostisch	x	✓	✓

Die nachfolgende Abb. 3.6 fasst die Arten der Erklärungen in einem Schaubild zusammen. Entsprechend der eingeführten Definition werden interpretierbare Modelle direkt auf den Trainingsdaten gelernt, sodass der Ansatz ausschließlich für die Dimension ante-hoc gilt. Im Allgemeinen ist ein interpretierbares Modell global, aber in den Fällen, in denen nur Teile der Daten einbezogen wurden, könnte es auch nur für die lokale Dimension gelten. Sofern ein interpretierbares Modell trainiert wird, wird stets ein spezifischer Ansatz zur Erzeugung der Erklärungen verwendet, der somit nicht agnostisch auf unterschiedliche Modellarchitekturen anwendbar ist. Surrogat-Modelle werden entweder parallel oder post-hoc generiert. Direkte Erklärungen benötigen

entweder ein interpretierbares oder ein Blackbox-Modell, somit ist dieser Ansatz nur post-hoc anwendbar. Darüber hinaus gibt es spezifische und agnostische Ansätze zur Generierung direkter Erklärungen, die die Eigenschaften des Modells entweder nutzen oder nicht.

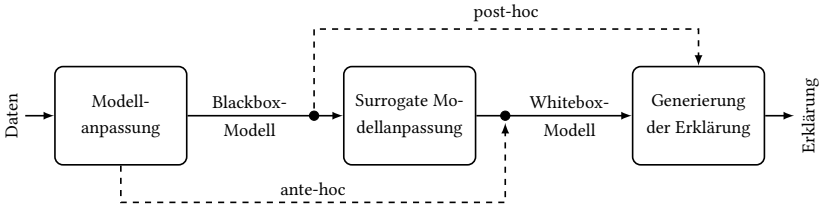


Abbildung 3.6: Zusammenfassung der unterschiedlichen Erklärarten [Quelle: In Anlehnung an [Bur21b], S. 257].

3.7 Zusammenfassung

In diesem Kapitel wurde das Vorgehen zur Extraktion unterschiedlicher Arten von Erklärungen vorgestellt. Insgesamt werden fünf Arten im Bereich des überwachten maschinellen Lernens unterschieden. Für jede Art von Erklärung wurde die zugrundeliegende Problemdefinition herausgearbeitet.

Wie das Kapitel aufgezeigt hat, werden *Whitebox-Modelle* zum einen in *inherente Whitebox-* und für die Erklärbarkeit zum anderen in *optimierte Whitebox-Modelle* unterschieden. Inherente Whitebox-Modelle sind auf natürliche Art und Weise für den Anwender nachvollziehbar. Optimierte Whitebox-Modelle werden gezielt dazu entworfen, um die Nachvollziehbarkeit der Modelle zu steigern. Dabei kann explizit zwischen der Genauigkeit als auch der Erklärbarkeit der Modelle variiert werden. *Surrogat-Modelle* sind interpretierbare Modelle, die basierend auf einem Blackbox-Modell gelernt werden. Diese können entweder global als Beschreibung für ein gesamtes Modell oder lokal für eine bestimmte Instanz extrahiert werden. *Direkte*

Erklärungen können entweder als lokale oder globale Erklärungen aus der Blackbox extrahiert werden.

Abschließend wurde die in Kapitel 2.2.6 eingeführte Taxonomie der Erklärbarkeit den unterschiedlichen Arten der Erklärungen zugeordnet. Bei Whitebox-Modellen handelt es sich ausschließlich um Ante-hoc-Verfahren, die zumeist global, jedoch auch lokal auf einer Teilmenge der Daten generiert werden können. Surrogat-Modelle werden ausschließlich post-hoc generiert und können sowohl lokal oder global als auch modell-spezifisch oder modell-agnostisch sein. Direkte Erklärungen können ausschließlich post-hoc, lokal oder global und modell-spezifisch oder modell-agnostsich sein. Die Erklärungen im Bereich der interpretierbaren Modelle bieten am wenigsten Flexibilität verglichen mit den beiden anderen Arten der Erklärungen. Modell-agnostische Verfahren ermöglichen die größte Flexibilität, da diese auf unterschiedliche Modelle angewendet werden können [Rib16b].

Bislang fehlte ein Vorgehensmodell zur Klassifizierung unterschiedlicher Arten von Erklärungen für den Bereich des überwachten maschinellen Lernens. Die Einführung des Vorgehensmodells schließt diese Lücke. Zusätzlich zeigt die Zuordnung relevanter Verfahren aus der Literatur für die jeweilige Erklärungsart die Anwendbarkeit des Vorgehensmodells zur Einordnung der Verfahren.

Im nachfolgenden Kapitel wird ein Verfahren entworfen, das ausgehend von einem Blackbox-Modell, die globale Erklärbarkeit durch ein Surrogat-Modell erzeugen soll. Diese Verfahren lässt sich ebenfalls in das Vorgehensmodell eingliedern.

4 Globale Surrogat-Modelle

In diesem Kapitel wird ein Verfahren entworfen, das dem Vorgehen globaler Surrogat-Modelle aus Abb. 3.1 (d) entspricht. Ausgehend von einem neuronalen Netz wird ein globales Surrogat-Modell in Form einer regelbasierten Liste erzeugt. Das Surrogat wird mithilfe der Regularisierung generiert, sodass dieses sowohl die Eigenschaften der Nachvollziehbarkeit als auch der Genauigkeit erfüllt. Das Ziel ist es, ein komplexes, tiefes, neuronales Netz für den Anwender nachvollziehbar zu gestalten. Die Arbeiten basieren auf den Veröffentlichungen von Burkart et al. [Bur19] sowie Burkart et al. [Bur20a]. Zunächst wird die betrachtete Problemstellung genauer definiert. Im Anschluss wird die allgemeine Methodik zur Lösung der Problemstellung vorgestellt. Darauf aufbauend werden relevante Vorarbeiten erläutert, die für die Optimierung des Verfahrens verwendet werden. In der Beschreibung der Zwischenergebnisse werden Probleme des vorgestellten Verfahrens mit möglichen Problemlösungen, die im Rahmen der Optimierung des Verfahrens umgesetzt wurden, betrachtet. Sowohl das Verfahren aus den Vorarbeiten als auch das optimierte Verfahren wird im Rahmen einer Evaluation untersucht und mit Referenzverfahren verglichen. Abschließend werden die gesammelten Resultate zusammengefasst.

4.1 Problemdefinition

Die Problemstellung betrachtet die binäre Klassifikation, die bereits in Abschnitt 2.1.2.1 eingeführt wurde. Für die Funktion $b : \mathcal{X} \rightarrow \mathcal{Y}$ wird für eine bestimmte Eingabe $x \in \mathcal{X}$ die Ausgabe $b(x) = y$ erlernt, sodass x zur Klasse y gehört. Da nur die binäre Klassifikation betrachtet wird, gilt für $\mathcal{Y} = \{0, 1\}$.

Die Funktion wird aus einem Datensatz der Dimension N gelernt, wie in Gleichung (4.1) dargestellt:

$$\mathcal{D} = \{(x_n, y_n) | x_n \in \mathcal{X}, y_n \in \mathcal{Y}, n \in \{1, \dots, N\}\}. \tag{4.1}$$

wobei die Merkmalsvektoren $x_n \in \mathcal{X}$ und die Klassenbezeichner $y_n \in Y$ sind. Das neuronale Netz b_θ ist ein mehrschichtiges Perzeptron (s. Abschnitt 2.5), das die Parameter θ, d. h. die Gewichte, aus einer Menge möglicher Parameter $\Theta \subseteq \mathbb{R}^n$ enthält. Die Verlustfunktion wurde bereits in Gleichung (2.1) definiert, die die Leistung der Klassifizierung bewertet. Das Ziel ist demnach die Suche nach optimalen Gewichten für θ, wie in Gleichung (4.2) dargestellt, sodass der Verlust minimal ist.

$$\theta^* = \arg\min_{\theta \in \Theta} L(\theta, \mathcal{D}). \tag{4.2}$$

Das resultierende Modell ist eine Blackbox. Die Ergebnisse der Blackbox sind für den Anwender ohne eine weitere Lösung nicht nachvollziehbar. Das Ziel ist es, die Blackbox nachvollziehbar zu gestalten, indem ein globales Surrogat-Modell erzeugt wird, das die Vorhersagen der Blackbox nachahmt. Globale Surrogat-Modelle wurden in Abschnitt 3.5.1 eingeführt und erläutert.

4.2 Methodik

In diesem Abschnitt wird das allgemeine Vorgehen beschrieben, durch das ein globales Surrogat-Modell mithilfe der Regularisierung generiert wird. Dieses allgemeine Vorgehen wird für das Verfahren in den Vorarbeiten und auch für das eigentliche Ergebnis des optimierten Verfahrens verwendet. Das resultierende globale Surrogat dient als nachvollziehbares Modell des neuronalen Netzes und wird mithilfe der Regularisierung in Form von Regellisten erzeugt. Die Methode der Regularisierung wurde bereits in Abschnitt 2.4 erläutert. Als

regelbasierte Surrogate werden Verfahren aus der Literatur verwendet, diese standen im Fokus von Abschnitt 2.5.2. Zunächst werden unter Abschnitt 4.2.2 relevante Vorarbeiten beschrieben, die zum Entwurf des optimierten Verfahrens geführt haben.

Erklärbarkeit muss nicht zwangsweise mit dem Verlust der Genauigkeit eines Modells einhergehen [Rud18]. Goodfellow et al. [Goo16] konstatieren, dass neuronale Netze oft gleich gute lokale Minima haben. Beispielsweise kann eine bestimmte Auswahl lokaler Minima nachvollziehbarer sein als eine andere. Die Grundidee der Vorgehensweise ist es, durch den Einsatz der Regularisierung während des Trainings eines neuronalen Netzes ein nachvollziehbares Surrogat-Modell in Form einer Regelliste zu trainieren. Aus diesem Surrogat-Modell wird eine Metrik zur Messung der Komplexität abgeleitet. Die Metrik wird als Strafterm auf die Zielfunktion des neuronalen Netzes addiert. Je komplexer das resultierende Surrogat ist, desto höher ist die Bestrafung. Die Regularisierung versucht, bei der Optimierung des Modells ein Fehlerminimum zu finden, um gleichzeitig ein genaues und zugleich nachvollziehbares Surrogat-Modell zu approximieren. Genauer gesagt fügt die Regularisierung einen Term Ω zur Verlustfunktion L hinzu. Dies führt zu der Zielfunktion, wie in Gleichung (2.12) des Abschnitts 2.4 bereits aufgezeigt wurde. Dabei ist $\lambda \in \mathbb{R}^{>0}$ die Stärke der Regularisierung und $\Omega(\cdot)$ die Länge der Regelliste. Die Regularisierung erlaubt, bestimmte Eigenschaften des Modells b_θ zu erzwingen. Um nachvollziehbare Modelle zu generieren, wird der Regularisierungsbegriff von einem regelbasierten Surrogat abgeleitet. Ein regelbasiertes Modell besteht aus einer Liste von Regeln $\mathcal{A}$. Die Regeln folgen der Form, wie in Gleichung (4.3) dargestellt.

$$\textbf{WENN } \mathrm{x}_n = a \ldots \textbf{ DANN } \mathrm{y}. \tag{4.3}$$

Eine solche Regel wird als Tupel (ρ, y) bezeichnet, wobei ρ eine Liste von k-Bedingungen und y die Vorhersage ist. Der Regularisierungsbegriff Ω bildet einen Zustand der Blackbox, gegeben durch die Gewichte auf die Komplexität eines entsprechenden regelbasierten Modells ab. Dazu wird das regelbasierte

Modell $w_{\mathcal{A}}$ unter Verwendung eines speziellen vorverarbeiteten Datensatzes $\mathcal{D}' = \{(x_n, b_\theta(x_n)) \mid x_n \in X, n \in \{1, \dots, N\}\}$ an die Vorhersagen der Blackbox angepasst.

Für das Training der regelbasierten Modelle werden die gleichen x_n aus $\mathcal{X}$ verwendet wie für das Training der Blackbox b_θ. Die Eingabedaten für die Regellisten unterliegen einer veränderten Vorverarbeitung, die im nächsten Abschnitt thematisiert wird.

4.2.1 Vorverarbeitung der Trainingsdaten

Das neuronale Netz erhält die Eingabedaten in numerischer Form aus $\mathbb{R}^n$. Kategorische Variablen können nicht als ganzzahlige Werte umgewandelt werden, da dies zu einem fehlerhaften Verhalten führen kann. Beispielsweise kann die Aktivierung für einen Wert höher sein als für einen anderen Wert. Sofern die Variablen untereinander eine bestimmte *Ordnung* haben, kann dies erwünscht sein. Jedoch stellt es bei Merkmalen, die keine Ordnung untereinander haben, ein fehlerhaftes Verhalten dar. Dabei können während des Trainings falsche Zusammenhänge erlernt werden. Daher werden die kategorischen Merkmale binär kodiert. Das heißt, aus einem Merkmal mit n möglichen Werten werden n Merkmale mit zwei möglichen Werten. Das i-te dieser n neuen Merkmale signalisiert, ob das vorliegende Objekt die Ausprägung i hat oder nicht.

Im Gegensatz dazu erwarten alle betrachteten Regellisten ihre Eingabe in Form von Kategorien. Zahlenwertige Merkmale können nicht ohne eine zielgenaue Vorverarbeitung verwendet werden. Aus diesem Grund wurde das sogenannte *Binning* angewendet. Bei diesem Verfahren werden zahlenwertige Merkmale in Kategorien eingeteilt.

4.2.2 Vorarbeiten

Das übergeordnete Vorgehen zum Training des regelbasierten Surrogates basiert auf den Arbeiten von Wu et al. [Wu18]. Anders als in diesen wurde anstatt eines Entscheidungsbaumes eine Regelliste verwendet und somit auch

ein veränderter Regularisierungsbegriff. Die nachfolgenden Abschnitte basieren auf den Arbeiten von Burkart et al. [Bur19] und Faller [Fal19]. Darin sind sowohl das Vorgehen als auch die Ergebnisse der Evaluation beschrieben. Das Verfahren wird nachfolgend teilweise auch als *RuleReg* bezeichnet.

4.2.2.1 Vorgehen

Zunächst wird als Regularisierungsbegriff die Menge aller Bedingungen ρ verwendet. Diese zählt die Gesamtzahl der vorhandenen Klauseln. Die Länge des regelbasierten Modells wird dementsprechend durch Gleichung (4.4) wie folgt berechnet:

$$\Omega_{w_A}(\theta) = \sum_{(\rho,y)\in\mathcal{A}} |\rho| \,. \tag{4.4}$$

quantifiziert.

Bei der Verwendung von Ω_{w_A} aus Gleichung (4.4) als Regularisierungsbegriff Ω in Gleichung (2.12) wird die Verlustfunktion L' nicht differenzierbar. Dies resultiert daraus, da das Training der Regelliste deren Länge bestimmt. Daher muss eine differenzierbare Proxy-Funktion wie in den Arbeiten von Wu et al. [Wu18] verwendet werden. Demnach wird zusätzlich ein neuronales Netz als Proxy q_ϕ verwendet, das die Vorhersagefunktion des neuronalen Netzes differenzierbar gestaltet. Demnach ist w_A ein Surrogat-Modell, das das Verhalten des Hauptmodells b_θ nachahmt, wobei q_ϕ eine Proxyfunktion bezeichnet, die sich Ω_{w_A} annähert. Mit der Proxyfunktion $q_\phi(\theta)$ wird die tatsächliche Verlustfunktion von b_θ damit wie folgt berechnet:

$$L''(\theta, \mathcal{D}) = L(\theta, \mathcal{D}) + \lambda \cdot q_\phi(\theta) \,. \tag{4.5}$$

Das Modell benötigt eine differenzierbare Zielfunktion $L'(\theta)$, da ein Gradientenabstiegsverfahren verwendet wird. Das *Proxy-Netz* q_ϕ erhält als Eingabe die kompletten Gewichte des Hauptmodells und lernt somit, die Länge der resultierenden Regelliste vorherzusagen. q_ϕ generiert folglich die Information für das Hauptmodell, wie die Länge einer surrogaten Regelliste sein wird.

Formal wird dabei der nachfolgende Datensatz:

$$\mathcal{D}_{\text{proxy}} = \{(\theta, \Omega_{w_A}(\theta)) \mid \theta \in \Theta\}\,. \tag{4.6}$$

erstellt und für das Training verwendet.

Für jedes θ wird ein regelbasiertes Modell w_A trainiert und dessen Komplexität gemessen. Schließlich wird q_ϕ auf dem Datensatz trainiert. Die Sammlung dieser θ muss nicht getrennt vom Training des Hauptmodells erfolgen. Der Proxy q_ϕ kann während des Trainings des Hauptmodells berechnet werden. Die Abtastrate und die Trainingsrate sind zwei Hyperparameter des Verfahrens. Bei jedem Abtastrate-Schritt des Gradientenabstiegs innerhalb des Trainings des Hauptmodells b_θ wird ein neuer Datenpunkt $(\theta, \Omega(\theta))$ zu D_{proxy} hinzugefügt. In jedem Trainingsraten-Gradientenschritt wird die Proxy-Funktion q_ϕ auf dem Datensatz $\mathcal{D}_{\text{proxy}}$ neu trainiert.

Die Dimension des Proxys q_ϕ hängt von der Anzahl der Parameter des Hauptmodells b_θ ab. Da das Hauptmodell ein neuronales Netz darstellt, muss die Eingabe x numerisch sein, d. h. $\mathcal{X} \subseteq \mathbb{R}^n$. Die verwendeten regelbasierten Modelle hingegen können nur kategorische Daten verarbeiten. Die Vorverarbeitung der Daten wurde bereits in Abschnitt 4.2.1 beschrieben. Für das regelbasierte Modell wird der nachfolgende Datensatz verwendet

$$\mathcal{D}' = \{(x'_n, b_\theta(y_n)) \mid x'_n \in \mathcal{X}', n \in \{1, \ldots, N\}\}\,, \tag{4.7}$$

wobei $x'_n \in \mathcal{X}'$ für die Verwendung mit regelbasierten Modellen vorverarbeitet wird.

4.2.2.2 Evaluierung

Das Verfahren wurde auf verschiedenen Datensätzen und unter Anwendung der in Abschnitt 2.5.2 beschriebenen regelbasierten Verfahren evaluiert. Als Referenzverfahren für die Evaluation wurden die Verfahren von Wu et al. [Wu18], die nativen regelbasierten Modelle, die direkt auf dem Datensatz trainiert werden, das nicht regularisierte neuronale Netz und ein nativer Entscheidungsbaum verwendet.

Nachfolgend werden zunächst die Datensätze und die verwendeten Hyperparameter beschrieben. Im Anschluss daran werden die Ergebnisse in Bezug auf die Laufzeit, die Komplexität und die Genauigkeit untersucht.

Datensätze

Insgesamt wurden vier Datensätze aus den Bereichen Krebsdiagnosen, Passagierdaten, Zensusdaten und Kreditbewilligungen verwendet. Der Einkommensdatensatz *Adult* stammt aus Zensusdaten der Vereinigten Staaten von Amerika[19]. Jeder Datensatz besteht aus Informationen über eine einzelne Person. Insgesamt gibt es 15 Merkmale, darunter sind beispielsweise die Merkmale *Alter*, *Bildung* oder *Heimatland*. Das Ziel ist es, basierend auf den Merkmalen einer Person vorherzusagen, ob die Person mehr als 50.000 US-Dollar pro Jahr verdient oder nicht.

Der Datensatz zur Krebsdiagnose *Cancer* besteht aus 32 Merkmalen, die Merkmale der Brustmasse beschreiben, die durch eine Feinnadelaspiration gewonnen wurden[20]. Die Merkmale enthalten Informationen wie den mittleren Radius, die Fläche oder die Glätte der vorhandenen Zellkerne. Das Ziel ist es, die Zellen als bös- oder gutartig zu klassifizieren.

Der Datensatz *Titanic* stellt die Passagierdaten der Titanic aus dem Jahr 1912 dar[21]. Insgesamt gibt es 11 Merkmale wie z. B. Alter, Passagierklasse oder Anzahl der Geschwister an Bord, die sowohl numerisch als auch kategorischer Art sind. Das Ziel ist es, vorherzusagen, ob ein Passagier die Katastrophe überlebt hat oder nicht.

Der Datensatz *FICO* befasst sich mit Kreditanträgen [Fic18]. Dieser wurde im Rahmen eines Wettbewerbs zur Thematik des erklärbaren maschinellen Lernens publiziert. Die Merkmale enthalten Informationen über die Kreditgeschichte und die Liquidität einzelner Kunden. Das Ziel ist es, zu klassifizieren, ob Kunden ihren Kredit in Zukunft tilgen können oder nicht.

[19] https://archive.ics.uci.edu/ml/datasets/adult, letzter Abruf am 26.05.2020.

[20] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), letzter Abruf am 26.05.2020.

[21] https://www.kaggle.com/c/titanic, letzter Abruf am 26.05.2020.

Tabelle 4.1: Hyperparameter der Blackbox-Modelle pro Datensatz [Quelle: In Anlehnung an [Bur19], S. 703].

	Titanic	Cancer	Adult	FICO
Epochen	50	30	50	60
Batch Size	40	10	250	100
Augmentation Samples	20	20	5	20
Proxy Epochen	600	300	9000	3000
Proxy Trainingsgrösse	50	10	20	20
Samplerate	1	1	5	1
Trainingsrate	4	5	40	4

Hyperparameter

Das Hauptmodell enthält zwei verborgene Schichten, die erste Schicht 25 Neuronen und die zweite Schicht 10 Neuronen. Als Verlustfunktion wird die *binäre Kreuzentropie* verwendet, die eine Wahrscheinlichkeitsverteilung zurückgibt. Für die Aktivierung aller verborgenen Schichten wurde *Relu* und *Sigmoid* für die Ausgabe des Hauptmodells verwendet. Die Hyperparameter der Proxy-Funktion wurden entsprechend dem Verfahren von Wu et al. [Wu18] gewählt. Für die verborgenen Schichten wurden 25 Neuronen, die Relu-Aktivierung sowie eine Softplus-Aktivierungsfunktion für die Ausgabe verwendet. Die Datensätze wurden jeweils in einen Trainings- (67 %) und einen Validierungsdatensatz (33 %) unterteilt. Tabelle 4.1 listet alle weiteren relevanten Hyperparameter auf.

Laufzeit

Die nachstehende Tabelle 4.2 zeigt die Laufzeiten der verschiedenen Verfahren. Die regelbasierten Surrogat-Modelle müssen häufig für ein Training des Hauptmodells trainiert werden. Unter den regelbasierten surrogat-Verfahren gilt das Verfahren *SBRL* als das schnellste. Das Verfahren zeigt weniger Varianz in der Laufzeit als die anderen regelbasierten Surrogate. Der Ansatz ist deutlich langsamer als das Referenzverfahren von Wu et al. [Wu18]. Dies liegt daran, dass Entscheidungsbäume effizienter aufgebaut werden können als Regellisten.

Tabelle 4.2: Durchschnittliche Laufzeiten der Verfahren [Quelle: In Anlehnung an [Bur19], S. 703].

Verfahren	Titanic	Cancer	Adult	FICO
RIPPER	$0{,}22 \pm 0{,}03$	$0{,}20 \pm 0{,}03$	$15{,}68 \pm 4{,}62$	$4{,}0 \pm 1{,}17$
CN2	$0{,}22 \pm 0{,}11$	$0{,}58 \pm 0{,}48$	$0{,}70 \pm 0{,}67$	$9{,}65 \pm 3{,}95$
SBRL	$0{,}04 \pm 0{,}014$	$0{,}12 \pm 0{,}03$	$0{,}96 \pm 0{,}06$	$1{,}37 \pm 0{,}17$
[Wu18]	$0{,}0 \pm 0{,}0$	$0{,}003 \pm 0{,}0$	$0{,}01 \pm 0{,}0$	$0{,}06 \pm 0.01$

Modelltreue

Die nachfolgende Tabelle 4.3 enthält die Ergebnisse der Modelltreue. Die regelbasierten Surrogate erzeugen eine hohe Modelltreue für fast alle Datensätze außer für den FICO-Datensatz.

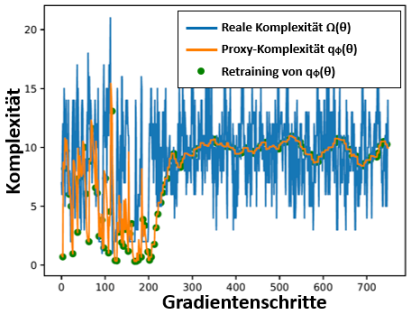
Tabelle 4.3: Modelltreue der Verfahren. Die besten Durchläufe für die jeweiligen Datensätze wurden *fett* hervorgehoben [Quelle: In Anlehnung an [Bur19]), S. 703].

Verfahren	Cancer	Titanic	Adult	FICO
RIPPER (regularisiert)	**0,9415**	**0,9729**	0,9897	0,8114
CN2 (regularisiert)	0,8989	0,7559	0,9324	0,6431
SBRL (regularisiert)	**0,9415**	0,9525	0,9585	**0,8268**
[Wu18]	0,9189	0,9424	**0,9997**	0,8097
SBRL Nativ	0.9096	0,9627	0,9870	0,8140

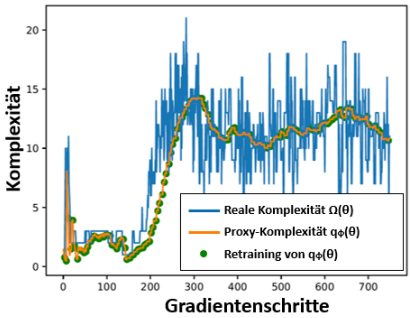
Einfluss der Regularisierungsstärke λ

Ein interessanter Aspekt bei der Regularisierung ist die Untersuchung des Einflusses von λ. Das Modell wurde mit unterschiedlichen Regularisierungsstärken λ getestet, um dies zu untersuchen. Unter der Annahme, dass die Nachvollziehbarkeit und die Genauigkeit der Modelle im Widerspruch zueinander stehen, ist es wünschenswert, beide Eigenschaften durch die Regularisierungsstärke λ gezielt kontrollieren zu können. Bei der Untersuchung

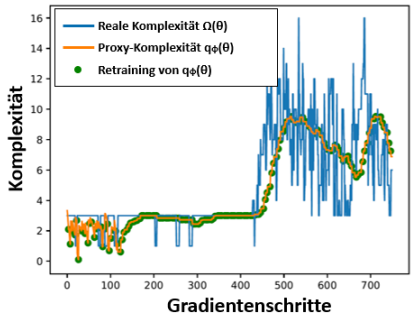
fällt auf, dass sich die Regularisierungsstärke während des Trainings stark auf die Komplexität der regelbasierten Modelle auswirkt. Das Hauptmodell gibt bei einer hohen Regularisierungsstärke lediglich einen Klassenbezeichner zurück, somit erzeugt das regelbasierte Modell nur eine Standardregel. Bei niedrigen Regularisierungsstärken verhält sich das Modell so, als würde keine Regularisierung angewendet. Demnach wurden viele Regeln durch das Surrogat erzeugt, sodass die Nachvollziehbarkeit nicht gegeben war. Die nachstehende Abb. 4.1 veranschaulicht die Ergebnisse für unterschiedliche Stärken der Regularisierung.



(a) $\lambda = 1$



(b) $\lambda = 100$



(c) $\lambda = 1000$

Abbildung 4.1: Ergebnisse für den Titanic-Datensatz (von oben nach unten $\lambda = 1, 100, 1000$) mit dem regelbasierten Surrogat SBRL [Quelle: In Anlehnung an [Bur19], S. 704].

Performance der Verfahren

Die Tabellen 4.4–4.7 zeigen die Ergebnisse für die einzelnen Datensätze, durch die ersichtlich wird, dass das Verfahren von Wu et al. [Wu18] genauer, wenngleich auch komplexer ist als alle regelbasierten Surrogate. Um die Werte der Regellisten und der Entscheidungsbäume besser vergleichen zu können, wurden für die Entscheidungsbäume die Anzahl der Entscheidungsknoten und die durchschnittliche Pfadlänge in Klammern dargestellt. Die Ergebnisse machen deutlich, dass die Surrogat-Modelle sowohl nachvollziehbar als auch modelltreu sind. Daher sollte ebenso untersucht werden, inwieweit sich die Regularisierung auf die Leistung des Hauptmodells auswirkt. Dazu wurde das Hauptmodell so trainiert, dass dieses mit einem nicht regularisierten Netz verglichen werden kann. Demnach wurden vor allem niedrige Regularisierungsstärken für λ verwendet und Läufe mit einer hohen Genauigkeit ausgewählt. In den meisten Fällen konnte die Genauigkeit des nicht regularisierten Modells erreicht werden. Bei einigen Datensätzen konnte eine hohe Genauigkeit nur mit langen, nicht nachvollziehbaren regelbasierten Modellen generiert werden. Diese waren insbesondere nicht kürzer als das Surrogat eines nicht regularisierten Modells. Bei dem Datensatz *Krebs* konnten die Surrogate die Genauigkeit der nicht regularisierten Modelle nicht erreichen. Der nativ trainierte Entscheidungsbaum führte zu genaueren Ergebnissen, jedoch war auch die Komplexität deutlich höher.

Tabelle 4.4: Ergebnisse für den Datensatz *Titanic* [Quelle: In Anlehnung an [Bur19], S. 704]. Bei den baumbasierten Verfahren steht die Anzahl der Regeln in Klammern.

	Verfahren	Genauigkeit	AUC	Komplexität
Titanic	RIPPER (regularisiert)	0,81	0,77	11
	RIPPER (nativ)	0,77	0,77	16
	CN2 (regularisiert)	0,75	0,70	23
	CN2 (nativ)	0,76	0,78	27
	SBRL (regularisiert)	0,80	0,77	11
	SBRL (nativ)	0,79	0,76	13
	[Wu18]	0,82	0,80	25 (13)
	Nicht-regularisiert (SBRL)	0,82	0,82	21
	Entscheidungsbaum (nativ)	0,83	0,82	248 (53)

Tabelle 4.5: Ergebnisse für den Datensatz *Cancer* [Quelle: In Anlehnung an [Bur19], S. 704].

	Verfahren	Genauigkeit	AUC	Komplexität
Cancer	RIPPER (regularisiert)	0,90	0,89	12
	RIPPER (nativ)	0,91	0,94	17
	CN2 (regularisiert)	0,78	0,81	11
	CN2 (nativ)	0,90	0,89	19
	SBRL (regularisiert)	0,89	0,89	4
	SBRL (nativ)	0,92	0,96	11
	[Wu18]	0,90	0,91	16 (6)
	Nicht-regularisiert (SBRL)	0,94	0,94	20
	Entscheidungsbaum (nativ)	0,95	0,94	41 (11)

Tabelle 4.6: Ergebnisse für den Datensatz *Adult* [Quelle: In Anlehnung an [Bur19], S. 704].

	Verfahren	Genauigkeit	AUC	Komplexität
Adult	RIPPER (regularisiert)	0,83	0,72	180
	RIPPER (nativ)	0,83	0,73	280
	CN2 (regularisiert)	0,83	0,72	93
	CN2 (nativ)	0,81	0,70	215
	SBRL (regularisiert)	0,82	0,72	32
	SBRL (nativ)	0,82	0,72	27
	[Wu18]	0,80	0,65	469 (85)
	Nicht-regularisiert (SBRL)	0,83	0,71	191
	Entscheidungsbaum (nativ)	0,86	0,76	4537 (699)

Tabelle 4.7: Ergebnisse für den Datensatz *FICO* [Quelle: In Anlehnung an [Bur19], S. 704].

	Verfahren	Genauigkeit	AUC	Komplexität
FICO	RIPPER (regularisiert)	0,72	0,71	73
	RIPPER (nativ)	0,52	0,50	1
	CN2 (regularisiert)	0,70	0,69	79
	CN2 (nativ)	0,70	0,70	167
	SBRL (regularisiert)	0,72	0,72	15
	SBRL (nativ)	0,71	0,70	28
	[Wu18]	0,71	0,71	613 (300)
	Nicht-regularisiert (SBRL)	0,70	0,70	68
	Entscheidungsbaum (nativ)	0,70	0,70	2364 (702)

Genauigkeit, Komplexität und Erklärbarkeit

Optimalerweise werden Modelle erzeugt, die sowohl genau als auch nachvollziehbar sind. Die nachstehenden Abb. 4.2-4.5 zeigen den AUC-Wert verschiedener Durchläufe in Bezug zur Komplexität. Bei allen Durchläufen erzeugten die regelbasierten Surrogate weniger komplexe Modelle als das Verfahren von Wu et al. [Wu18]. Somit ist es möglich, das Hauptmodell beim Training in einer Weise zu optimieren, dass ohne größere Verluste bei der Genauigkeit

nachvollziehbare Surrogate erstellt werden können. Daraus lässt sich die Tendenz ableiten, dass die Verfahren *RIPPER* und *CN2* komplexere regelbasierte Modelle erzeugen als SBRL.

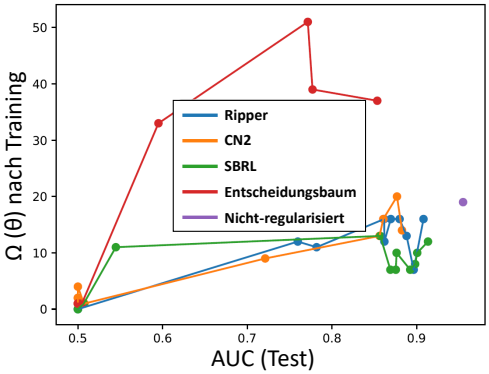


Abbildung 4.2: Datensatz *Cancer* [Quelle: In Anlehnung an [Bur19], S. 705].

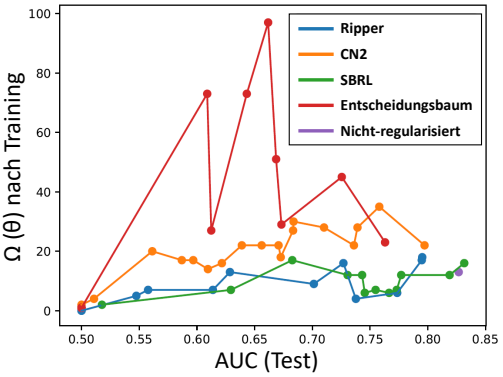


Abbildung 4.3: Datensatz *Titanic* [Quelle: In Anlehnung an [Bur19], S. 705].

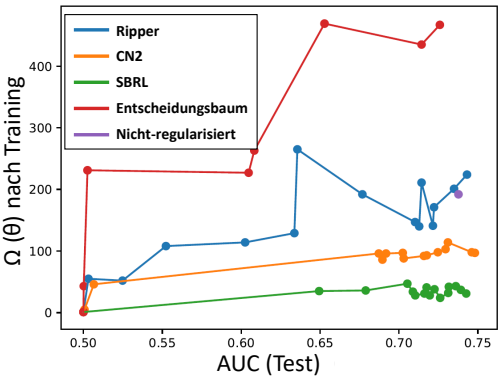


Abbildung 4.4: Datensatz *Adult* [Quelle: In Anlehnung an [Bur19], S. 705].

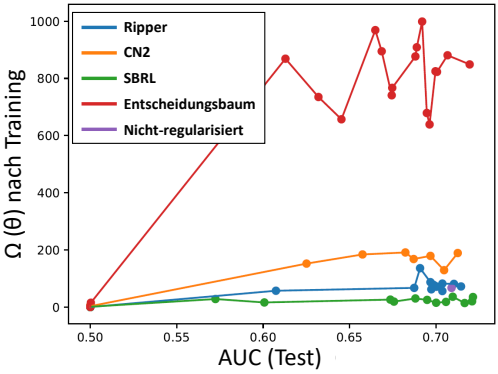


Abbildung 4.5: Datensatz *FICO* [Quelle: In Anlehnung an [Bur19], S. 705].

Insgesamt erzeugten die regelbasierten Surrogate weniger als 20 Regeln für die Datensätze *Titanic* und *Krebs*. Die Ergebnisse für die Datensätze des UCI Credit und der FICO waren deutlich komplexer. Das nachfolgende Listing 4.1 zeigt ein Beispiel einer surrogaten Regelliste für den Datensatz *Titanic*.

```
WENN Passagierklasse = 1 UND Geschlecht = weiblich DANN
    Überlebenschance = Lebend
WENN Geschlecht = weiblich UND IstAlleine = Nicht alleine
    DANN Überlebenschance = Lebend
WENN Titel = Miss DANN Überlebenschance = Lebend
WENN Geschlecht = weiblich DANN Überlebenschance = Tot
WENN Passagierklasse = 3 DANN Überlebenschance = Tot
WENN Fare = (-0.512, 51.233] UND IsAlone = Alone DANN
    Überlebenschance = Tot
WENN Passagierklasse = 1 DANN Überlebenschance = Lebend
WENN IstAlleine = Nicht Alleine DANN Überlebenschance = Tot
WENN  Überlebenschance = Tot
```

Listing 4.1: Beispiel einer Regelliste für den Datensatz *Titanic* [Quelle: In Anlehnung an [Bur19], S. 703].

4.2.2.3 Diskussion der Zwischenergebnisse

Ein Nachteil des beschriebenen Verfahrens ist, dass für das Training des Hauptmodells nach jedem Gradientenschritt ein neues regelbasiertes Surrogat trainiert werden muss und dadurch die Laufzeit stark ansteigt. Zusätzlich muss auch der Proxy mehrfach nachtrainiert werden. Das Verfahren benötigt somit mehr Zeit zum Trainieren verglichen mit den nicht regularisierten Verfahren. Die Ergebnisse der Evaluation zeigen, dass sich das Verfahren *SBRL* am besten als regelbasiertes Surrogat eignet. Über mehrere Durchläufe und verschiedene Datensätze hinweg produziert das Verfahren sowohl genaue als auch nachvollziehbare Regellisten mit einer hohen Modelltreue.

Eine weitere Problematik der regelbasierten Surrogate zeigt sich in einer hohen Varianz der Ergebnisse. Diese unterscheiden sich insbesondere stark in ihrer Klassifikationsleistung und der Länge des dazugehörigen Surrogates. Aus Sicht des Anwenders verringert sich dadurch die Verlässlichkeit eines Systems.

Für einige Datensätze – bspw. bei *Adult* (siehe Tabelle 4.6) – erzeugen native Regellisten bereits nachvollziehbare Regellisten mit einer hohen Genauigkeit. Dies bestätigt die Aussage von Rudin [Rud18], dass nicht für jeden Zweck eine Blackbox trainiert werden muss, um eine hohe Genauigkeit zu erreichen. Die Ergebnisse zeigen jedoch auch (siehe Tabelle 4.7), dass das regelbasierte Surrogat *SBRL* eine hohe Genauigkeit bei nahezu halbierter Komplexität erreicht.

4.2.3 Idee der Optimierung auf Basis der Vorarbeiten

Die zuvor in Abschnitt 4.2.2.3 beschriebenen Probleme des Verfahrens werden in diesem Kapitel aufgegriffen und optimiert. Das Ziel ist es, wie zu der

allgemeinen Methodik unter Abschnitt 4.2 erläutert, ein tiefes neuronales Netz zu regularisieren, sodass ein globales Surrogat-Modell angepasst werden kann, das aus einer nachvollziehbaren Regelliste besteht. Die nachstehenden Abb. 4.6-4.7 zeigen die Vorgehen der beiden Verfahren im Einzelnen. Das Verfahren, das nachfolgenden entworfen wird, wird auch als *GiniReg* bezeichnet.

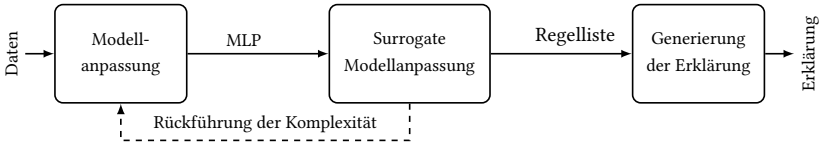


Abbildung 4.6: Vorgehen des Verfahrens aus den Vorarbeiten (RuleReg) [Quelle: In Anlehnung an [Bur19], S. 703].

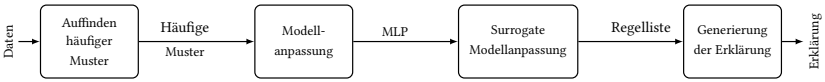


Abbildung 4.7: Vorgehen des optimierten Verfahrens (GiniReg) [Quelle: In Anlehnung an [Bur19], S. 703].

Im Gegensatz zu dem in Abschnitt 4.2.2 beschriebenen Verfahren soll die Regularisierung so gestaltet sein, dass diese differenzierbar ist und eine gradientenbasierte Optimierung ohne ein Proxy-Modell ermöglicht. Das Vorgehen der beiden Verfahren ist in der nachstehenden Abb. 4.6 für das Verfahren *RuleReg* und in Abb. 4.7 für das Verfahren *GiniReg* dargestellt. Die Hauptunterschiede liegen darin, dass das optimierte Verfahren basierend auf dem Auffinden von häufigen Mustern arbeitet und im Gegensatz zu *RuleReg* nicht nachtrainiert werden muss. Um dieses Vorgehen umzusetzen, wird ein sog. stapelweises (engl. batch-wise) Vorgehen der Regularisierung eingeführt. Die Kernidee ist, dass häufig gemeinsam auftretende Muster (sog.

Mini-Batches) in den Trainingsdaten denselben Klassenbezeichner besitzen. Diese *Mini-Batches* sind zufällig zusammengesetzt; ein Minimum für diesen Fehler anzunehmen ist, allen denselben Klassenbezeichner zu geben. Demnach wird eine Strafe für unterschiedliche Klassenbezeichner vergeben, wenn Objekte sich leicht durch eine Regel zusammenfassen lassen würden bzw. wenn diese zu einem *Cluster* gehören. Um die Laufzeit für das Training zu verkürzen, wird das Surrogat nur auf den Mini-Batches trainiert. Das Training des Surrogates auf den Mini-Batches wird bestraft, je nachdem wie unterschiedlich die Vorhersagen der Klassenbezeichner auf dem Mini-Batch sind.

4.2.3.1 Auffinden häufig auftretender Muster

In Abschnitt 4.2.2 wurde das Verfahren *SBRL* als regelbasiertes Surrogat verwendet. Der Algorithmus besteht aus zwei wesentlichen Elementen. Zunächst werden häufige Muster gesucht, bevor im Anschluss diesen Mustern Klassenbezeichner zugewiesen werden. Häufige Muster bilden in einer Regel die Antezedenzien ab und werden wie in Gleichung (4.8) abgebildet.

$$\mathrm{x}_n = \alpha \textbf{ UND } \mathrm{x}_m < \beta. \tag{4.8}$$

Der *Support* einer häufig auftretenden Regel bzw. eines Musters $\mathrm{supp}(\mathcal{A})$ (s. Gleichung (4.9)) wird als die Menge aller Datenpunkte beschrieben, die dieses Muster erfüllen.

$$\mathrm{supp}(\mathcal{A}) = \{(x, \cdot) \in \mathcal{D} \mid \forall a \in \mathcal{A} : a(x) = \mathit{Wahr}\}. \tag{4.9}$$

Ein Muster wird als *häufiges Muster* bezeichnet, wie in nachfolgender Formel dargestellt, wenn ein bestimmter Anteil des Datensatzes dieses Muster erfüllt bzw. einen bestimmten Schwellwert ϵ erreicht.

$$\frac{1}{|\mathcal{D}|}|\text{supp}(\mathcal{A})| \geq \epsilon. \tag{4.10}$$

Jede Teilmenge des Musters ist dabei auch ein häufiges Muster. Die Elemente des Supports supp($\mathcal{A}$) werden als *Items* bezeichnet und meinen die Elemente der Menge. Die supp($\mathcal{A}$) wird als *Itemset* bezeichnet.

4.2.3.2 Vorgehen

In den Vorarbeiten wurde aufgezeigt (s. Abschnitt 4.2.2), dass das Verfahren *SBRL* als regelbasiertes Surrogat am geeignetsten ist. Daher wird bei der Optimierung des Verfahrens nur dieses als Verfahren zur Erzeugung des regelbasierten Surrogates betrachtet. *SBRL* generiert die Regeln, indem den häufigsten Mustern Klassenbezeichner zugeteilt werden. Die Paare aus Antezedenzien und Klassenbezeichnern bilden die Regeln. Die Idee ist, dass SBRL weniger Regeln benötigt, um den gesamten Datensatz abzudecken, wenn alle Objekte, die einem Muster angehören, denselben Klassenbezeichner besitzen. Der Grundgedanke ist somit, dass häufig auftretende Merkmalsausprägungen in Regeln zusammengefasst werden.

Angenommen, es liegt ein Datensatz vor, in dem vier häufige Muster mit dem gleichen Klassenbezeichner auftreten, würden dabei vier Regeln entstehen. Im Gegensatz dazu würde ein Muster, das viele Datenpunkte mit unterschiedlichen Klassenbezeichnern umfasst, viele unterschiedliche Regeln ergeben. Daher besteht das Ziel darin, häufige Muster, die die gleichen Klassenbezeichner besitzen, aufzufinden, um so die Länge der resultierenden surrogaten Regelliste zu reduzieren.

4.2.3.3 Der Regularisierungsbegriff

Analog zu dem Verfahren, das in den Zwischenergebnissen beschrieben wurde, wird auch bei der Optimierung des Verfahrens Ω_{w_A} auf die Zielfunktion des Hauptmodells addiert, um dieses entsprechend regularisieren zu können. Demnach wird versucht nachfolgende Gleichung, entsprechend dem Ω in Gleichung (2.12), zu optimieren:

$$L''(\theta, \mathcal{D}) = L(\theta, \mathcal{D}) + \lambda \cdot \Omega_{w_A}(\theta). \tag{4.11}$$

Um $\Omega_{w_A}(\theta)$ zu erhalten, werden vor dem Training des Hauptmodells die häufigsten Muster im Datensatz gesucht. Das Resultat dessen ist eine Liste $\mathcal{F}$, die alle häufigen Muster enthält. Für ein häufiges Muster $F \in \mathcal{F}$ sollen homogene Klassenbezeichner gefunden werden. Dabei wird die nachfolgende Gleichung minimiert:

$$\frac{1}{2 \cdot F} \sum_{i,j \in \mathcal{F}} y_i - y_j, \tag{4.12}$$

wobei $y_i = b_\theta(x_i)$, also die Ausgabe des Hauptmodells, für den i-ten Datenpunkt ist. Der Strafterm Ω_{w_A} wird, wie in der nachstehenden Gleichung berechnet:

$$\begin{aligned} \Omega_{w_A} &\triangleq \frac{1}{2 \cdot F \cdot N} \sum_{F \in \mathcal{F}} \sum_{i,j \in F} |y_i - y_j| \\ &\leq \frac{1}{2 \cdot F} \sum_{F \in \mathcal{F}} \frac{1}{F} \sum_{i,j \in F} |y_i - y_j| \\ &\leq \frac{1}{2 \cdot F} \sum_{F \in \mathcal{F}} \sum_{i,j \in F} |y_i - y_j|. \end{aligned} \tag{4.13}$$

Um Ω_{w_A} effizient berechnen zu können, werden die folgenden Schritte durchgeführt. Sei $o \triangleq (1, \dots, 1)^{\mathrm{T}} \in \mathbb{R}^N$ und $\hat{y} \triangleq (\hat{y}_1, \dots, \hat{y}_N)^{\mathrm{T}}$

$$\mathbf{Y} \triangleq o \cdot y^{\mathrm{T}} = \begin{pmatrix} y^{\mathrm{T}} \\ \vdots \\ \hat{y}^{\mathrm{T}} \end{pmatrix} \in \mathbb{R}^{N2}. \tag{4.14}$$

und

$$\mathbf{V} \triangleq \mathbf{Y} - \mathbf{Y}^{\mathrm{T}} = \begin{pmatrix} |y_1 - y_1| & |y_2 - y_1| & \dots \\ |y_1 - y_2| & |y_2 - y_2| & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}, \tag{4.15}$$

ist die Matrix der paarweisen Differenz zwischen den Ausgaben $V_{i,j} = y_j - y_i$. Um die Summe der Differenz im gleichen Cluster F zusammenzufassen, wird die Matrix $\mathbf{R}^F \in \mathbb{R}^{N \times N}$ mit folgenden Elementen definiert:

$$\mathbf{R}^F_{i,j} \triangleq \begin{cases} 1, & x_i, x_j \in \mathcal{F} \\ 0, & \text{alle anderen} \end{cases}. \tag{4.16}$$

Wird diese elementweise multipliziert, folgt daraus

$$\mathbf{V} \circ \mathbf{R}^F = \begin{pmatrix} |y_1 - y_1| \cdot R^F_{1,1} & |y_2 - y_1| \cdot R^F_{2,1} & \dots \\ |y_1 - y_2| \cdot R^F_{1,2} & |y_2 - y_2| \cdot R^F_{2,2} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}, \tag{4.17}$$

eine Matrix mit allen Differenzen zwischen den Elementen in F. Daraus folgt der Regularisierungsbegriff mit:

$$\begin{aligned} \Omega_{w_A} &= \frac{1}{2 \cdot |\mathcal{F}| \cdot N} \sum_{i,j \in F} \sum_{F \in \mathcal{F}} \left(\mathbf{V} \circ \mathbf{R}^F \right)_{i,j} \\ &= \frac{1}{2 \cdot |\mathcal{F}| \cdot N} \sum_{i,j \in F} \left(\mathbf{V} \circ \sum_{F \in \mathcal{F}} \mathbf{R}^F \right)_{i,j}. \end{aligned} \tag{4.18}$$

Weiter wird die Matrix $\mathbf{C} \in \mathbb{R}^{N \times |\mathcal{F}|}$ definiert als

$$C_{i,j} \triangleq \begin{cases} 1, & x_i \in \mathcal{F}_j \\ 0, & \text{für alle anderen} \end{cases} . \tag{4.19}$$

Diese enthält eine 1 an der Position i, j, wenn die Instanz x_i im Itemset F_j enthalten ist und ansonten eine 0. Für einzelne Zeilen aus C ergibt sich daraus:

$$C_{:,j} \cdot C_{:,j}^T = \begin{pmatrix} C_{1,j} \cdot C_{1,j} & C_{1,j} \cdot C_{2,j} & \dots \\ C_{2,j} \cdot C_{1,j} & C_{2,j} \cdot C_{2,j} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} = \mathbf{R}^{F_j}$$

und somit,

$$\mathbf{C} \cdot \mathbf{C}^{\mathrm{T}} = \sum_{F \in \mathcal{F}} \mathbf{R}^F . \tag{4.20}$$

Der Regularisierungsbegriff kann somit definiert werden als:

$$\Omega_{w_A} = \frac{1}{2 \cdot |\mathcal{F}| \cdot N} \sum_{1 \le i,j \le N} \left(\mathbf{V} \circ (\mathbf{C} \cdot \mathbf{C}^{\mathrm{T}}) \right)_{i,j} . \tag{4.21}$$

Für die Berechnung von Ω_{w_A} aus Gleichung (4.18) werden alle Elemente der Matrix $\mathbb{V} \circ (\mathbf{C} \cdot \mathbf{C}^{\mathrm{T}})$ summiert. Diese Formulierung ist differenzierbar, weshalb kein Proxy-Modell benötigt wird. Die Matrix C wird dem Modell separat übergeben. Die Berechnung der Matrix wird über das *Frequent Item Mining* des Verfahrens *SBRL* ausgeführt.

4.2.3.4 Evaluierung

Das optimierte Verfahren wurde im Rahmen einer Evaluation näher untersucht. Dieses wurde mit dem Verfahren aus den Zwischenergebnissen, dem Verfahren von Wu et al. [Wu19] (s. Abschnitt 2.4.2) und dem nicht regulierten neuronalen Netz verglichen.

Datensätze

Neben den Datensätzen aus Abschnitt 4.2.2.2 wurden vier weitere verwendet. Der Datensatz *UCI Bank Marketing* enthält Daten zu Marketingkampagnen von Banken. Dabei werden Empfänger unterschieden, die sich basierend auf einer Marketingkampagne gemeldet haben, und solche, die sich nicht gemeldet haben. Der Datensatz *MAGIC Gamma Telescope* besteht aus simulierten Daten von hochenergetischen Gamma-Teilchen eines Teleskops. Die Klassenbezeichner wurden in *Rauschen* und *Signal* unterteilt. Der Datensatz *Wine* enthält Informationen über Weine aus Portugal. Jede Instanz beschreibt die Eigenschaften eines Weins. Der Wein wurde zwischen 0 (schlecht) und 10 (gut) bewertet. Die Bewertungen wurden im Rahmen der binären Klassifikation in zwei Klassen aufgeteilt. Weine mit einer Bewertung von fünf und höher wurden als gut und kleiner fünf als schlecht gekennzeichnet.

Hyperparameter

Das Hauptmodell b_θ ist bei allen Experimenten ein mehrschichtiges Perzeptron. Als Verlustfunktion wurde die binäre Kreuzentropie verwendet. Die Aktivierungsfunktion für alle versteckten Schichten ist *ReLU* und für die Ausgabeschicht *sigmoid*. Als Optimierungsalgorithmus wurde *Adam* eingesetzt. Die Anzahl der *Items* kann Tabelle 4.9 entnommen werden. Der Datensatz wurde randomisiert in einen Trainings- (67 %) und einen Validierungsdatensatz (33 %) aufgeteilt. In der nachstehenden Tabelle 4.8 sind alle weiteren relevanten Hyperparameter aufgelistet.

Tabelle 4.8: Hyperparameter der MLPs pro Datensatz [Quelle: In Anlehnung an [Bur19], S. 5].

Datensatz	Epochen	Batch-Größe	Neuronen	Lernrate
Bank	500	128	[512, 256]	1e-4
Wine	500	128	[512, 256]	1e-4
Gamma	500	128	[512, 256]	1e-4
Adult	500	128	[512, 256]	1e-4
Cancer	30	15	[15, 10]	4e-3
Fico	500	128	[512, 256]	1e-4
Titanic	400	64	[16, 16, 8, 8]	4e-1
Toy	100	8	[12, 4]	3e-1

Tabelle 4.9: Anzahl häufiger Items [Quelle: [Bur19], S. 5].

	Gamma	Wine	Titanic	Cancer	Bank	Adult	Fico
$\lvert\mathcal{F}\rvert$	1099	1354	651	32197	3778	2117	9264

Einfluss der Regularisierungsstärke

Die nachfolgende Abb. 4.8 zeigt die Ergebnisse der Modelle, die mit verschiedenen Regularisierungsstärken λ trainiert wurden. Für jede Regularisierungsstärke wurden 20 Modelle mit identischen Hyperparametern trainiert und die Standardabweichung der Modelle gemessen. Dabei wird ersichtlich, dass mit steigendem λ die Anzahl der Regeln innerhalb der Surrogate sinkt. Somit kann die Komplexität der Surrogate reduziert werden. Die Standardabweichung der Surrogate steigt für höhere Regularisierungsstärken nicht wesentlich an.

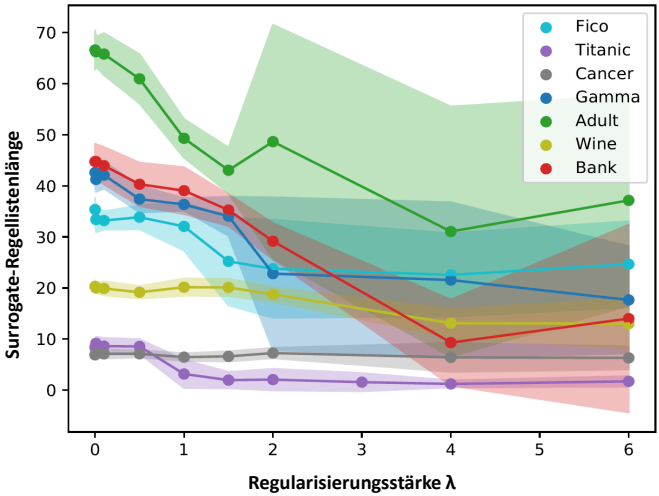


Abbildung 4.8: Komplexität des Surrogates bei verschiedenen Regularisierungsstärken λ. Für ein steigendes λ sinkt die Komplexität der Listen [Quelle: In Anlehnung an [Bur19], S. 5].

Modelltreue

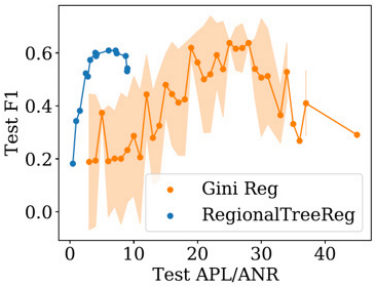
Bei allen ausgeführten Durchläufen zeigten die Surrogate eine hohe Modelltreue gegenüber dem Hauptmodell. Die nachstehende Tabelle 4.10 zeigt die durchschnittliche Modelltreue über alle Läufe und Regularisierungsstärken λ auf. In diesem Zusammenhang wird deutlich, dass die Werte der Modelltreue zwischen 80 % und 95 % liegen. Daraus kann abgeleitet werden, dass die Surrogate modelltreue Erklärungen erzeugen.

Tabelle 4.10: Modelltreue der Verfahren pro Datensatz [Quelle: In Anlehnung an [Bur19], S. 6].

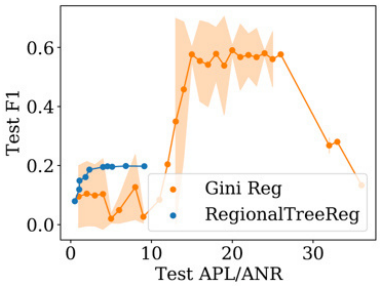
Datensatz	GiniReg	[Wu19]	RuleReg
Bank	$0{,}94 \pm 0{,}05$	0,89	-
Wine	$0{,}92 \pm 0{,}06$	0,88	-
Gamma	$0{,}92 \pm 0{,}05$	0,88	-
Cancer	$0{,}82 \pm 0{,}11$	-	0,94
Fico	$0{,}84 \pm 0{,}08$	-	0,83
Titanic	$0{,}98 \pm 0{,}03$	-	0,95
Adult	$0{,}98 \pm 0{,}01$	0,91	0,96

Erklärbarkeit

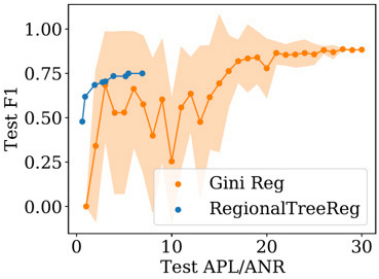
In Abb. 4.9 wird das vorgestellte Verfahren GiniReg mit den Ergebnissen des Verfahrens von Wu et al. [Wu19] verglichen. Wie bereits in den Vorarbeiten erwähnt, ist der Begriff der durchschnittlichen Entscheidungspfadlänge nicht direkt auf Regellisten übertragbar. Daher wird die durchschnittliche Anzahl derjenigen Regeln verwendet, die zur Klassifizierung einer Stichprobe herangezogen werden. Die Darstellungen verdeutlichen, dass das vorgestellte Verfahren genauere Modelle ermöglicht als das Verfahren von Wu et al. [Wu19]. Für Surrogate mit ähnlicher Komplexität ist das Verfahren von Wu et al. [Wu19] etwas genauer als das entworfene Verfahren. Letzteres zeichnet sich bei hoher Regularisierungsstärke durch mehr Unsicherheit in der resultierenden Länge der Regellisten aus. Listing 4.2 zeigt eine resultierende Regelliste.



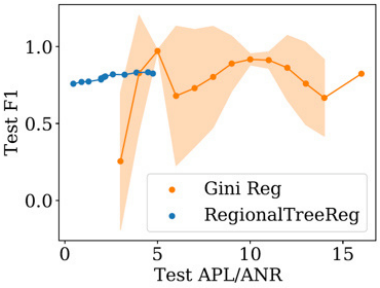
(a) Datensatz *Adult*



(b) Datensatz *Bank*



(c) Datensatz *Magic Gamma Telescope*



(d) Datensatz *Wine*

Abbildung 4.9: Durchschnittliche Pfadlänge (APL, für Regional TreeReg)/Durchschnittliche Anzahl von Regeln (ANR, für GiniReg) des resultierenden Surrogat-Modells aufgetragen gegen F_1-Score auf den unterschiedlichen Datensätzen. GiniReg generiert genauere Modelle im Vergleich zum Referenzverfahren von Wu et al. [Wu19] [Quelle: In Anlehnung an [Bur19], S. 7].

```
WENN (Fläche in (181.1, 592.1]) DANN Diagnose = gutartig
SONST (Glätte in (90.6, 110.6]) UND
SONST (Konkavität  in (-0.0004, 0.04]) DANN Diagnose = gutartig
SONST (konkave Punkte in (0.09, 0.1]) UND
SONST(Konkavität  in (0.04, 0.09]) DANN Diagnose = gutartig
SONST (Textur in (15.8, 19.5]) UND
SONST(Konkavität in (0.04, 0.09]) DANN Diagnose = gutartig
SONST (Umfang in (13.6, 16.4]) DANN Diagnose = gutartig
SONST Diagnose = bösartig
```

Listing 4.2: Beispiel einer Regelliste für den Datensatz *Cancer* [Quelle: In Anlehnung an [Bur19], S. 6].

Laufzeit

In Tabelle 4.11 wird die Laufzeit der unterschiedlichen Modelle verglichen. Dabei ist zu erkennen, dass das optimierte Verfahren schneller ist als das Verfahren aus den Vorarbeiten, da das Surrogat-Modell einmal trainiert wird.

Tabelle 4.11: Laufzeit in Sekunden [Quelle: In Anlehnung an [Bur20a] S. 6].

Datensatz	GiniReg	SBRL Nativ	RuleReg
Bank	3827,88 ±108,36	117,16 ±8,86	-
Wine	231,08 ±13,91	6,23 ±0,68	-
Gamma	897,67 ±58,11	60,0 ±5,17	-
Cancer	28,37 ±8,36	2,88 ±0,18	369,81 ±52,19
Fico	1541,49 ±42,74	21,38 ±1,61	21755,96 ±212,03
Titanic	78,17 ±21,31	2,51 ±0,24	662,67 ±16,70
Adult	1936,76 ±90,41	97,31 ±6,2	4513,51 ±147,3

Performance der Verfahren

Über die Regularisierungsstärke λ wurde eine Rastersuche durchgeführt und das Modell mit höchsten F_1-Werten ausgewählt. Bei der Rastersuche wird nach optimalen Hyperparametern gesucht. Bei jeder Regularisierungsstärke wurden 20 zufällige Durchläufe durchgeführt. Diese wurden mit dem akkuratesten nicht regularisierten Modell ($\lambda = 0$) aus 20 Durchläufen und der akkuratesten nativen SBRL-Regeliste (ebenfalls aus 20 Durchläufen) verglichen. Die Ergebnisse sind in Tabelle 4.12 zusammengefasst. Für das optimierte Verfahren beziehen sich die Spalten F_1 und die Genauigkeit auf die Leistung des Hauptmodells. Bei allen Datensätzen ist das optimierte Verfahren genauer als das nicht regulierte neuronale Netz. Dies ist ein Indikator dafür, dass das entworfene Verfahren nicht nur die Länge des resultierenden Surrogates reduziert, sondern auch das Hauptmodell regularisiert, sodass zugleich die Generalisierung des Hauptmodells verbessert wird. Auf nahezu allen Datensätzen (ausgenommen dem Datensatz *Wine*) ist das optimierte Verfahren genauer als das nativ trainierte Verfahren *SBRL*. Da der Datensatz *Wine* unausgewogene Klassenbezeichner enthält, ist es wahrscheinlich, dass das nativ trainierte Verfahren *SBRL*, bedingt durch den Bayes'schen Ansatz, die Klassen einer Minderheit besser verarbeitet als das auf dem neuronalen Netz optimierte Verfahren. Die nachstehende Tabelle 4.12 zeigt, dass das optimierte Verfahren genauere Modelle, jedoch auch eine etwas höhere Komplexität des resultierenden Surrogates erzeugt.

Tabelle 4.12: Performancewerte der Verfahren [Quelle: In Anlehnung an [Bur19], S. 7].

	Verfahren	Hauptmodell F_1	Hauptmodell ACC	Surrogat F_1	Surrogat ACC	APL/ANR	Modelltreue	λ
Cancer	GiniReg	**0,96** ± 0,0	0,97 ± 0,0	0,85 ± 0,02	0,9 ± 0,01	4,0 ± 0,53	0,9 ± 0,02	0,5
	NeuralNet	**0,96** ± 0,01	0,97 ± 0,0	0,84 ± 0,02	0,9 ± 0,01	3,96 ± 0,53	0,9 ± 0,02	0,0
	SBRL	0,84 ± 0,02	0,9 ± 0,01	-	-	3,4 ± 0,36	-	-
Fico	GiniReg	**0,7** ± 0,0	0,72 ± 0,0	0,69 ± 0,0	0,71 ± 0,0	13,98 ± 1,32	0,89 ± 0,0	0,01
	NeuralNet	**0,7** ± 0,0	0,72 ± 0,0	0,69 ± 0,0	0,71 ± 0,0	14,62 ± 1,32	0,89 ± 0,0	0,0
	SBRL	0,68 ± 0,01	0,7 ± 0,0	-	-	6,81 ± 0,85	-	-
Gamma	GiniReg	**0,9** ± 0,0	0,86 ± 0,0	0,87 ± 0,0	0,81 ± 0,0	17,64 ± 1,33	0,91 ± 0,0	1,0
	NeuralNet	0,88 ± 0,0	0,85 ± 0,0	0,85 ± 0,0	0,81 ± 0,0	25,6 ± 1,33	0,86 ± 0,0	0,0
	SBRL	0,87 ± 0,0	0,81 ± 0,0	-	-	17,64 ± 1,72	-	-
	Wu et al. [Wu19]	0,75	-	-	-	6,91	-	-
Titanic	GiniReg	**0,74** ± 0,01	0,78 ± 0,01	0,71 ± 0,01	0,77 ± 0,01	5,27 ± 1,56	0,95 ± 0,02	0,01
	NeuralNet	**0,74** ± 0,01	0,78 ± 0,01	0,71 ± 0,01	0,77 ± 0,01	5,47 ± 1,56	0,96 ± 0,02	0,0
	SBRL	0,7 ± 0,0	0,76 ± 0,0	-	-	5,04 ± 0,3	-	-
Wine	GiniReg	0,97 ± 0,0	0,94 ± 0,01	0,97 ± 0,0	0,94 ± 0,01	6,67 ± 2,25	0,98 ± 0,01	4,0
	NeuralNet	0,88 ± 0,0	0,8 ± 0,01	0,91 ± 0,0	0,83 ± 0,01	10,42 ± 2,25	0,84 ± 0,01	0,0
	SBRL	**0,98** ± 0,0	0,96 ± 0,0	-	-	5,96 ± 0,14	-	-
	Wu et al. [Wu19]	0,83	-	-	-	4,49	-	-
Adult	GiniReg	**0,65** ± 0,0	0,8 ± 0,01	0,65 ± 0,0	0,81 ± 0,01	24,42 ± 2,06	0,98 ± 0,0	0,5
	NeuralNet	0,64 ± 0,0	0,78 ± 0,01	0,64 ± 0,0	0,78 ± 0,01	25,67 ± 2,06	0,98 ± 0,0	0,0
	SBRL	0,61 ± 0,01	0,83 ± 0,0	-	-	15,48 ± 1,01	-	-
	Wu et al. [Wu19]	0,61	-	-	-	6,06	-	-
Bank	GiniReg	**0,62** ± 0,0	0,89 ± 0,0	0,53 ± 0,01	0,88 ± 0,0	19,35 ± 3,02	0,91 ± 0,0	1,0
	NeuralNet	0,58 ± 0,0	0,85 ± 0,0	0,53 ± 0,01	0,88 ± 0,0	21,95 ± 3,02	0,87 ± 0,0	0,0
	SBRL	0,43 ± 0,02	0,9 ± 0,0	-	-	17,56 ± 2,2	-	-
	Wu et al. [Wu19]	0,2	-	-	-	6,83	-	-

4.2.3.5 Diskussion der Ergebnisse

Das optimierte Verfahren generiert verkürzte Regellisten und liefert dabei weiterhin eine hohe Genauigkeit. Mit zunehmender Regularisierungsstärke λ nimmt die Länge der surrogaten Regelliste ab. Im Vergleich zu dem Verfahren von Wu et al. [Wu19] erzeugt das optimierte Verfahren Regellisten, die länger sind, jedoch im Gegenzug eine höhere Genauigkeit aufweisen. Außerdem zeigt das Surrogat eine hohe Modelltreue gegenüber dem Hauptmodell. Auch das Trainieren des optimierten Verfahrens ist deutlich schneller als das Verfahren aus den Vorarbeiten. Die Ergebnisse des optimierten Verfahrens zeigen einen weiteren gewünschten Effekt, nämlich eine niedrigere Varianz innerhalb der Ergebnisse. So werden bei einer Neuberechnung der Modelle keine stark abweichenden Modelle generiert. Abschließend lässt sich erkennen, dass das optimierte Verfahren die Probleme des Verfahrens aus den Zwischenergebnissen behebt und mit den verglichenen Referenzverfahren konkurrieren kann.

4.3 Zusammenfassung

In diesem Kapitel wurde zunächst ein allgemeines Vorgehen beschrieben, das basierend auf dem Konzept der Regularisierung ein globales Surrogat erzeugt. Dieses Surrogat entwirft eine nachvollziehbare Regelliste, die als Interpretation für ein neuronales Netz dienen soll. Innerhalb des Abschnitts Zwischenergebnisse wurde ein Verfahren beschrieben, das in Anlehnung an das von Wu et al. [Wu18] anstatt eines Entscheidungsbaumes nachvollziehbare Regellisten erzeugt. Als Regularisierungsterm wird die Anzahl der Antezedenzien verwendet. Je länger also die Regelliste ist, desto höher fällt die Bestrafung aus. Das Verfahren wurde auf unterschiedlichen Datensätzen evaluiert.

Das Basisverfahren aus den Zwischenergebnissen reduzierte zwar die Komplexität der Regelliste, allerdings wurde auch ein Rückgang der Klassifikationsleistung beobachtet. Die erzeugten Surrogate erreichten eine Modelltreue

von 90 % gegenüber dem Hauptmodell. Da das Verfahren einige Schwachstellen aufzeigte, wurde ein optimiertes Verfahren entworfen. Bei der Optimierung war es das Ziel, die Schwachstellen des Basisverfahrens aus Wu et al. [Wu18], wie z. B. die Laufzeit, zu korrigieren. Bedingt dadurch, dass das Surrogat nach jedem Gradientenschritt neu trainiert wird, stieg die Laufzeit zur Berechnung einer Lösung stark an. Das optimierte Verfahren erzeugt das Surrogat nur noch wenige Male. Dabei wird eine Strafe für unterschiedliche Klassenbezeichner vergeben, wenn Objekte sich leicht durch eine Regel zusammenfassen lassen bzw. wenn sie zu einem *Cluster* gehören. Ein weiteres Problem stellte die nicht direkte Differenzierbarkeit der Zielfunktion dar. Zunächst musste ein neuronales Netz als Proxy verwendet werden, um die Zielfunktion differenzierbar zu machen. Der Proxy diente dazu, die Länge der Regelliste zu schätzen. Im Zuge der Optimierung wurde nun eine Matrix aus der Menge der häufigsten Muster erstellt, die direkt in jedem Gradientenschritt aus der Vorhersage des Hauptmodells berechnet werden kann. Somit entfällt die Verwendung des neuronalen Netzes als Proxy.

Die Länge der resultierenden Regelliste verkürzte sich bei dem optimierten Verfahren. Mit zunehmender Regularisierungsstärke λ nimmt die Länge des Surrogates ab. Im Vergleich zu dem Verfahren von Wu et al. [Wu19] sind die resultierenden Listen geringfügig länger, wohingegen die Modelle auch zu genaueren Ergebnisse führten. Die Modelltreue erreichte bei fast allen getesteten Datensätzen bessere Ergebnisse. Somit stimmen die Vorhersagen der Surrogate vorwiegend mit den Vorhersagen der Hauptmodelle überein. Die Laufzeit der Modelle verkürzte sich stark und wurde durch die Reduktion der Anzahl an Trainingsdurchläufen der Surrogate optimiert. Die Ergebnisse ergaben zudem deutlich weniger Varianz als das Verfahren aus den Zwischenergebnissen.

Nachdem in diesem Kapitel ein Verfahren entworfen wurde, das ein globales Surrogat-Modell erzeugt, soll im nachfolgenden Kapitel betrachtet werden, wie Erklärungen für einzelne Instanzen erzeugt werden können. In einigen Fällen ist es nicht notwendig ein gesamtes Modellverständnis zu erzeugen sondern es reicht ein lokale Erklärung für eine entsprechende Instanz zu erzeugen.

5 Lokale Surrogat-Modelle

In diesem Kapitel wird ein Verfahren entworfen, das dem Vorgehen zur Extraktion lokaler Surrogat-Modelle aus Abb. 3.1 (f) entspricht. Auf Basis eines neuronalen Netzes wird ein lokales Surrogat-Modell erzeugt. Das Surrogat wird unter Verwendung einer kontrafaktischen Instanz und weiterer Stützpunkte aus der näheren Umgebung erstellt. Das Ziel dessen ist es, eine Erklärung zu erzeugen, die nahe an einer Entscheidungsgrenze des Modells liegt. Im Gegensatz zu Kapitel 4 wird dabei die lokale Erklärbarkeit anstatt der globalen Erklärbarkeit erzeugt. Die Inhalte des Kapitels wurden in Burkart et al. [Bur21a] veröffentlicht. Nachfolgend wird zunächst die betrachtete Problemstellung erläutert. Im Anschluss wird die Methodik des entworfenen Verfahrens zur Lösung der Problemstellung vorgestellt. Darauf aufbauend wird das entworfene Verfahren im Rahmen einer Evaluation mit einem Referenzverfahren untersucht und verglichen. Abschließend werden die Ergebnisse zusammengefasst.

5.1 Problemdefinition

Die vorliegende Problemstellung ist, wie auch in Kapitel 4, die binäre Klassifikation. Die Klassifikation im Allgemeinen wurde in Abschnitt 2.1.2.1 eingeführt. So wird für die Funktion $b : \mathcal{X} \rightarrow \mathcal{Y}$ eine bestimmte Eingabe $x \in \mathcal{X}$ die Ausgabe $b(x) = y$ erlernt, sodass x zur Klasse y gehört. Auch hier gilt für die binäre Klassifikation $\mathcal{Y} = \{0, 1\}$. Um die Vorhersage für die Instanz x mit $b(x) = y$ zu erklären, muss zunächst eine kontrafaktische Instanz x' mit $b(x') = y'$ gefunden werden. Ausgehend von der ersten kontrafaktischen Instanz soll die nächstgelegene Entscheidungsgrenze gefunden werden, die

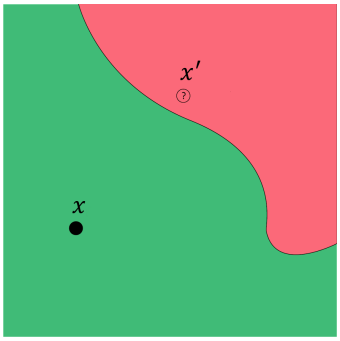
zwischen x und x' liegt. Die Entscheidungsgrenze soll mit einem Modell, dem lokalen Surrogat, approximiert werden, um die Entscheidung nachvollziehen zu können. Verfahren wie LIME [Rib16b], die lokale Surrogat-Modelle erstellen, die Erklärungen in der Umgebung der zu erklärenden Instanz x erzeugen, veranschaulichen, welche Merkmale zur Vorhersage einer bestimmten Instanz beigetragen haben. LIME erzeugt die Modelle in der Umgebung der zu erklärenden Instanz und nicht zwangsweise in der Umgebung einer naheliegenden Entscheidungsgrenze. Somit ist keine umfassende Nachvollziehbarkeit dessen gegeben, wie bspw. eine Entscheidung eines Modells verändert werden kann. Daher ist die Idee, die Nachvollziehbarkeit an einer lokalen Entscheidungsgrenze zu erzeugen. Dadurch soll es für den Anwender möglich sein, die Entscheidung an der lokalen Entscheidungsgrenze näher zu inspizieren. Diese Art der Erklärung soll dem Anwender neben der nachvollziehbaren Entscheidung auch die Möglichkeit geben, die Entscheidung des Modells zielgerichtet verändern zu können. Die Darstellung der Erklärung, die das gesamte lokale Surrogat-Modell abbildet, unterstützt den Anwender dabei, den Entscheidungsprozess simulieren zu können.

5.2 Methodik

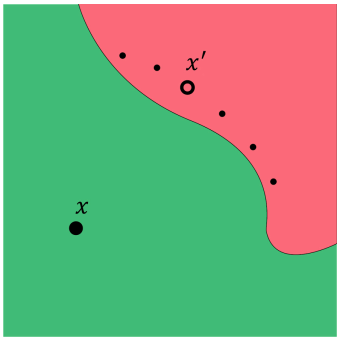
Nachfolgend wird das Vorgehen zur Lösung der vorangegangenen Problemdefinition, dem Auffinden einer Erklärung, die in der Umgebung einer Entscheidungsgrenze liegt, erläutert. Die Lösung unterteilt sich in ein phasenweises Vorgehen, das insgesamt fünf Phasen umfasst. Die erste Phase befasst sich mit der Suche nach einer kontrafaktischen Instanz, die in der Nähe zur ursprünglichen Instanz liegt. Ausgehend von dieser kontrafaktischen Instanz werden in der zweiten Phase Stützpunkte gesucht. Stützpunkte sind weitere kontrafaktische Instanzen, die in der Umgebung der ersten kontrafaktischen Instanz liegen. Basierend auf der ersten kontrafaktischen Instanz und den weiteren Stützpunkten wird in der dritten Phase die Entscheidungsgrenze gesucht. Nachdem diese gefunden wurde, wird in der vierten Phase ein lokales Surrogat in der Umgebung trainiert. Dieses wird dazu verwendet, um die Entscheidungen in der Umgebung der Entscheidungsgrenze nachvollziehen

zu können. Die fünfte und letzte Phase befasst sich mit der Darstellung der Erklärung. Dabei wird zwischen vier unterschiedlichen Arten der Darstellung unterschieden.

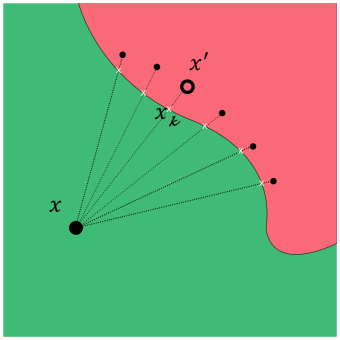
In der nachstehenden Abb. 5.1 ist das Vorgehen der einzelnen Phasen näher dargestellt. In den nachfolgenden Abschnitten (s. Abschnitt 5.2.1-5.2.5) werden die einzelnen Phasen genauer erläutert.



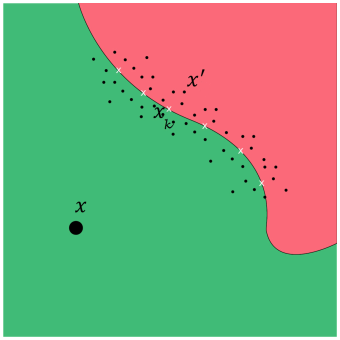
(a) Phase 1



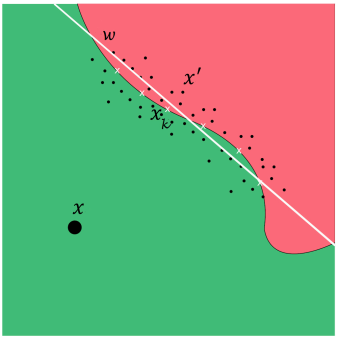
(b) Phase 2



(c) Phase 3



(d) Phase 4



(e) Phase 5

Abbildung 5.1: Phasenweises Vorgehen zum Auffinden eines lokalen Surrogates an einer Entscheidungsgrenze [Quelle: In Anlehnung an [Bur21a], S. 85].

5.2.1 Phase 1 – Suche nach kontrafaktischer Instanz

Die Suche nach einer kontrafaktischen Instanz bildet den ersten Schritt des phasenweisen Vorgehens ab. Wachter et al. [Wac17] stellen ein Vorgehen vor, mit dem diese aufgefunden werden kann. Eine kontrafaktische Instanz der faktischen Instanz x ist eine Instanz x', die ein verändertes Ergebnis erzeugt. Daher wird nach $b(x) \neq b(x')$ gesucht, wobei $b : \mathcal{X} \rightarrow \{0,1\}$ ein binäre Klassifikator ist. Idealerweise ist x' *nahe* an x im Merkmalsraum von $\mathcal{X}$ gemessen mit der Distanzmetrik $d(x, x')$. Die Distanzmetrik formalisiert, dass die kontrafaktische Instanz möglichst nahe an der ursprünglichen Instanz x ist. Die Kernidee von Wachter et al. [Wac17] besteht darin, die Suche nach einer kontrafaktischen Instanz als Optimierungsproblem zu formulieren. Formal wird die nachfolgende Verlustfunktion definiert:

$$L(x, x', y', \lambda) = \lambda \cdot \left(b(x') - y'\right)^2 + (1 - \lambda) \cdot d(x, x') \tag{5.1}$$

optimiert. In dieser entspricht y' dem veränderten Ergebnis der kontrafaktischen Instanz für x', und λ ist ein Parameter zwischen 0 und 1, der die Klasse der kontrafaktischen Instanz x' mit dem Abstand zur ursprünglichen Instanz x gewichtet. Der Abstand wird durch die nachstehende Gleichung gemessen:

$$d(x, x') = \frac{1}{n} \sum_{i}^{n} \frac{|x_i - x'_i|}{\text{MAD}_i} \ , \tag{5.2}$$

wobei n die Dimension des Merkmalraums und x_i das i-te Merkmal der Instanz x angibt. Die MAD entspricht der mittleren absoluten Abweichung und berechnet sich durch

$$\text{MAD}_i = \text{median}_{x \in D} \left(x_i - \bar{x}_i\right) \ , \tag{5.3}$$

mit $\bar{x}_i = \text{median}_{x \in D}(x_i)$.

für eine Instanz x des Trainingsdatensatz D. Die kontrafaktische Instanz wird gezielt durch das Lösen des nachstehenden Optimierungsproblems

$$L(x, y', \lambda) = \arg\min_{x'} \lambda \cdot \left(b(x') - y'\right)^2 + d(x, x') \tag{5.4}$$

gefunden. Zur Lösung der Gleichung kann bspw. das Optimierungsverfahren *Nelder-Mead* [Nel65] angewendet werden. Allerdings können auch weitere Methoden, wie z. B. das *ADAM*-Verfahren, sofern die Gradienten des Modells verwendet werden können, zum Einsatz kommen.

5.2.2 Phase 2 – Suche nach Stützpunkten

Das Ziel der zweiten Phase ist es, durch iterative Bewegungen innerhalb eines bestimmten Radius der ersten kontrafaktischen Instanz weitere Stützpunkte zu finden, die auch kontrafaktische Instanzen sind. Pro Iteration erweitert sich der Radius, diese Erweiterung wird als *Sektor* bezeichnet. Aus jedem Sektor wird zufällig eine bestimmte Anzahl von Punkten entnommen. Anhand der Ergebnisse wird geprüft, ob die Suche weiter in diese Richtung stattfinden wird oder ob ein neuer Sektor generiert werden soll. Sobald ein bestimmter Anteil der entnommenen Punkte nicht kontrafaktisch ist, wird die Suche in diese Richtung abgebrochen. Gegeben ist die Instanz x und die kontrafaktische Instanz x', die den ersten Stützpunkt der Menge I bildet. Darauf basierend wird nach weiteren Stützpunkten $I \subseteq \mathcal{X}$ gesucht. Die Stützpunkte sind alle kontrafaktische Instanzen.

Um sicherzustellen, dass geeignete Stützpunkte gefunden werden, wird die Suche in einem Umkreis $\mathcal{U}(r_i, r_o) = \{z \in \mathcal{X} \mid r_i \leq \|z\| \leq r_o\}$ eingeschränkt. Dabei wird $r_i = d - l$ und $r_o = d + l$ festgelegt, wobei d wie zuvor definiert und l die Breite des Sektors ist.

Zu diesem Zweck wird $d \coloneqq x' - x$ zwischen der ursprünglichen Instanz und der kontrafaktischen Instanz aus Phase 1 bestimmt. Dazu wird eine Stichprobe von Stützpunkten x'_i, $i > 1$ entnommen, indem d um x gedreht wird. Dabei

werden Punkte mit dem Abstand $||d||$ von x entnommen, die in der Nähe von x' liegen. Dies entspricht der Entnahme der Stützpunkte aus der Menge,

$$\mathcal{I}(x, x', a) = \{z \in \mathcal{X} : ||z-x|| = ||x'-x|| \text{ und } ||z-x'|| \leq a \text{ und } b(z) = y'\}, \tag{5.5}$$

wobei $||.||$ die euklidische Norm ist.

Die Suche nach Stützpunkten ist beendet, sobald Instanzen nicht mehr als kontrafaktisch klassifiziert werden. Die Einschränkung, dass nur Instanzen um x' verwendet werden, soll dazu führen, eine zusammenhängende Entscheidungsgrenze aufzufinden. Hierbei werden andere mögliche Entscheidungsgrenzen vernachlässigt, sodass eine möglichst einfache Erklärung gefunden wird [Mol20].

5.2.3 Phase 3 – Suche nach lokaler Entscheidungsgrenze

Ausgehend von den Stützpunkten I, die Resultat der Phase 2 sind, soll die Entscheidungsgrenze gefunden werden. Diese soll sich auf dem Segment zwischen x und $x'_i \in \mathcal{I}$ befinden. Das Segment mit $J_i(v) := x + v \cdot (x'_i - x)$ mit $v \in [0,1]$. Das Ergebnis dieser Phase ist eine abstrakte Darstellung der Entscheidungsgrenze in lokaler Nähe zu x. Dies kann eine Menge von Punkten K sein, sodass jedes $x_k \in K$ auf der Entscheidungsgrenze liegt. Diese Punkte werden auch Kontaktpunkte [Lau18] genannt. Aufgrund des Vorgehens zum Auffinden der Stützpunkte wird angenommen, dass sich der Wert von $\hat{b}$ monoton auf dem Segment J_i entwickelt. Dies gilt unter Umständen nicht für die Vorhersage von b. Unter dieser Annahme wird die binäre Suche auf dem Segment verwendet, um $x_k = J_i(v)$ für v zu lokalisieren und dadurch die Anzahl der Abfragen an das Modell b von $\mathcal{O}(n)$ auf $\mathcal{O}(\log(n))$ zu reduzieren.

5.2.4 Phase 4 – Trainieren des lokalen Surrogat-Modells

Basierend auf dem Ergebnis der Phase 3 wird eine Stichprobe von Instanzen T um die Entscheidungsgrenze entnommen. Auf dieser Stichprobe T wird ein nachvollziehbares Modell w als Surrogat trainiert, um die Entscheidungsgrenze lokal zu approximieren. Ähnlich wie bei dem Verfahren LIME [Rib16b] wird die Komplexität $\Omega(w)$ des Modells eingeschränkt, indem bspw. die maximale Tiefe für Entscheidungsbäume oder die Anzahl der Gewichte für lineare Modelle festgelegt wird. Formal wird w aus einer Menge von nachvollziehbaren Whitebox-Modellen $\mathcal{W}$ wie z. B. nicht tiefen Entscheidungsbäumen oder linearen Modellen generiert durch

$$L(T, b, L') = \arg\min_{w \in \mathcal{W}} \sum_{x \in T} L'\big(b(x), w(x)\big)\,, \tag{5.6}$$

wobei L' die Verlustfunktion mit z. B. dem mittleren quadratischen Fehler ist.

Das Rahmenwerk erlaubt es, die Anzahl der Merkmale, die vom Surrogat w berücksichtigt werden, zu begrenzen. Sofern kein Vorwissen über die konkrete Auswahl von Merkmalen vorhanden ist, können Methoden aus der Literatur wie *LARS* [Efr04] zur Einschränkung des Merkmalraumes verwendet werden.

5.2.5 Phase 5 – Darstellung der Erklärung

Der letzte Schritt besteht darin, dem Anwender die gefundenen Erklärungen entsprechend zu präsentieren. Mit dem Ergebnis aus Phase 4 können unterschiedliche Arten von Erklärungen generiert werden. Insgesamt können vier unterschiedliche Arten von Erklärungen dargestellt werden:

Darstellung der Merkmalsrelevanz: Die Merkmalsrelevanz kann für die ursprüngliche Instanz, die Unterstützungspunkte oder die kontrafaktische Instanz dargestellt werden. Wie Ribeiro et al. [Rib16a] in ihren Arbeiten festgestellt haben, kann die Relevanz eines Merkmals eine effiziente Möglichkeit sein, eine Entscheidung post hoc zu verstehen.

Darstellung der relativen Differenz: Mit der Darstellung der relativen Differenz zwischen der faktischen und der kontrafaktischen Instanz kann der Anwender direkte Handlungen ableiten.

Darstellung des lokalen Surrogates: Für den Aspekt der *Simulierbarkeit* (s. Abschnitt 2.2.5) ist es wünschenswert, dem Anwender eine gesamte Darstellung des lokalen Modells zu liefern. Dazu wird bspw. ein Entscheidungsbaum mit der Tiefe 2 oder 3 verwendet.

Darstellung über vordefinierte Textbausteine: Die Erklärungen können durch die Verwendung vordefinierter Textbausteine leicht aufbereitet und dargestellt werden.

Die Merkmalsrelevanz kann bspw. wie bei *LIME* dargestellt werden, sodass für eine gegebene Instanz die Merkmale extrahiert werden, die für die Vorhersage eines Modells am ausschlaggebendsten sind. Neben der Darstellung der Merkmalsrelevanz für die ursprüngliche Instanz kann auch die Merkmalsrelevanz der kontrafaktischen Instanz dargestellt werden.

Zwischen beiden Instanzen kann die relative Differenz dargestellt werden, die den Unterschied zwischen der faktischen und der kontrafaktischen Instanz angibt. Dieser Unterschied kann dem Anwender dabei helfen, gezielte Maßnahmen zu ergreifen, um eine gegebene Entscheidung zu verändern. Um dem Anwender die Handlungsfähigkeit zu ermöglichen, ist es wichtig, ihm aufzuzeigen, welche Merkmale geändert werden müssen, um eine gewünschte Vorhersage zu erreichen.

Die Darstellung des kompletten Surrogat-Modells soll den Anwender bei der Simulierbarkeit [Lip18] (s. Abschnitt 2.2.5) unterstützen. Die Simulierbarkeit ermöglicht es dem Anwender, den Entscheidungsprozess des Modells so zu

verstehen, dass dieser das Verhalten des Modells simulieren kann. Das Surrogat kann dank der reduzierten Komplexität vollständig von dem Anwender verstanden werden.

Die Verwendung vordefinierter Textbausteine eignet sich auch speziell dafür, Anwendergruppen ohne technischen Hintergrund Ergebnisse der Modelle möglichst nachvollziehbar zu vermitteln. Sowohl faktische als auch kontrafaktische Erklärungen eignen sich, um in Text ausgedrückt zu werden. Ausgehend vom ursprünglichen Modell b, der Instanz x und der kontrafaktischen Instanz x' können die Erklärungen wie folgt aufgebaut sein:

- Die Vorhersage für x war $b(x)$. Wenn x jedoch x' gewesen wäre, wäre die Vorhersage $b(x')$ gewesen.
- Um die Vorhersage von $b(x)$ auf $b(x')$ zu ändern, müssen die Werte $y - x$ der Merkmale verändert werden.

5.3 Evaluation

Das entworfene Verfahren wird in diesem Abschnitt evaluiert. Es wurde mit zwei unterschiedlichen Arten als lokales Surrogat trainiert: einem Entscheidungsbaum und einem linearen Modell. Als Referenzverfahren aus der Literatur wurde zum Vergleich das Verfahren LIME [Rib16a] verwendet. Nachfolgend werden der Aufbau und die Ergebnisse der Evaluation beschrieben.

5.3.1 Verwendete Datensätze

Als Datensätze wurden die wie auch bei der Evaluierung für das Verfahren aus Kapitel 4 verwendet. Daneben wurden noch zwei weitere Datensätze hinzugezogen. Der Datensatz *KDD* [Tav09] und der Datensatz *IDS* [Sha18]. Beide genannten Datensätze enthalten Informationen zu Netzwerkangriffen.

5.3.2 Hyperparameter

Für alle Experimente werden identische Hyperparameter verwendet. Das Modell b_θ ist bei allen Experimenten ein mehrschichtiges Perzeptron. Dieses enthält jeweils zwei verborgene Schichten – die erste Schicht besteht aus zwanzig und die zweite Schicht aus fünf Neuronen. Als Verlustfunktion wird die *binäre Kreuzentropie* eingesetzt. Für die Aktivierung aller verborgenen Schichten der Ausgabe wird *Relu* verwendet. Das neuronale Netz optimiert den mittleren quadratischen Fehler mit dem Optimierungsalgorithmus *Adam* bei einer Lernrate von 0,001. Der Datensatz wurde zufällig in einen Trainings- (67 %) und einen Validierungsdatensatz (33 %) unterteilt.

5.3.3 Modellperformance und Modelltreue

Zunächst wurden die Modelle b_θ trainiert und entsprechend ihrer Performance bewertet. Die Ergebnisse für die jeweiligen Datensätze sind in der nachstehenden Tabelle 5.1 dargestellt. Die resultierenden Modelle für die Datensätze *KDD* und *IDS* erreichten insgesamt eine hohe Genauigkeit. Alle trainierten Modelle der weiteren Datensätze bewirkten eine geringere Genauigkeit. Da das Augenmerk der Arbeit nicht auf der Optimierung der Modelle an sich, sondern auf der Nachvollziehbarkeit liegt, wurde keine weitere Optimierung der Modelle zur Leistungssteigerung vorgenommen.

Tabelle 5.1: Performance der Hauptmodelle [Quelle: In Anlehnung an [Bur21a], S. 88].

	Genauigkeit	F1
KDD	0,99	0,99
IDS	0,99	0,99
FICO	0,71	0,71
Credit	0,81	0,68
Titanic	0,91	0,87
Wine	0,97	0,94

Die Ergebnisse der Modelltreue wurden unter Verwendung der zehnfachen Kreuzvalidierung generiert und sind in Tabelle 5.2 dargestellt. Als Surrogate wurden zum einen ein lineares Modell und zum anderen ein Entscheidungsbaum trainiert. Die Repräsentation der Entscheidungsgrenze ist maßgeblich von den gewählten Stützpunkten abhängig, die in Phase 2 gefunden werden. Dementsprechend können sich die resultierenden Modelle unterscheiden; auch die Qualität der Surrogate variiert.

Tabelle 5.2: Modelltreue auf unterschiedlichen Datensätzen [Quelle: In Anlehnung an [Bur21a], S. 88].

	Lineares Modell	Entscheidungsbaum	LIME
KDD	0,87	0,97	0,85
IDS	0,96	0,99	0,93
FICO	0,96	0,97	0,86
Credit	0,99	0,99	0,95
Titanic	0,86	0,79	0,78
Wine	0,91	0,79	0,69

Die Ergebnisse machen deutlich, dass die lokalen Surrogate (lineares Modell und Entscheidungsbaum) eine höhere Modelltreue erreichen, verglichen mit dem Referenzverfahren LIME [Rib16b]. Das lokale Surrogat, das als Entscheidungsbaum trainiert wurde, generiert im Vergleich zum lokalen Surrogat, das als lineares Modell trainiert wurde, öfter eine höhere Modelltreue. Je nach Komplexität der lokalen Entscheidungsgrenze kann diese besser durch einen Entscheidungsbaum dargestellt werden.

5.3.4 Visualisierung der einzelnen Phasen am Beispiel des Kreditdatensatzes

In diesem Abschnitt werden die einzelnen Phasen anhand eines Beispiels visualisiert. Die nachstehende Abb. 5.3 zeigt die Ergebnisse der einzelnen Phasen am Datensatz *Credit*. Die initial ausgewählte Instanz mit dem Buchstaben (I) ist in Blau hervorgehoben. Alle weiteren Instanzen, die die gleiche Zielvariable wie die Instanz besitzen, sind ebenfalls Blau markiert. Im Rahmen der ersten Phase wird die erste kontrafaktische Instanz gefunden – diese ist mit dem Buchstaben (C) in Orange markiert. In der zweiten Abbildung ist das Resultat der Phase 2 dargestellt. Deren Ziel ist es, Stützpunkte aufzufinden, die in der Nähe der ersten kontrafaktischen Instanz liegen und ebenso kontrafaktisch sind. In Phase 3 wird die binäre Suche zwischen der ursprünglichen Instanz und den Stützpunkten durchgeführt. Dabei werden Punkte gefunden, die direkt auf der Entscheidungsgrenze liegen. Diese Punkte zeigen eine erste Visualisierung der Entscheidungsgrenze. In Phase 4 werden Stichproben für das Training des lokalen Surrogates entnommen, um das lokale Surrogat zu trainieren. Dabei ist die Entscheidungsgrenze deutlich sichtbar. In Phase 5 wird die Darstellung der Erklärung erzeugt, die im nächsten Abschnitt an einem konkreten Beispiel näher erläutert wird.

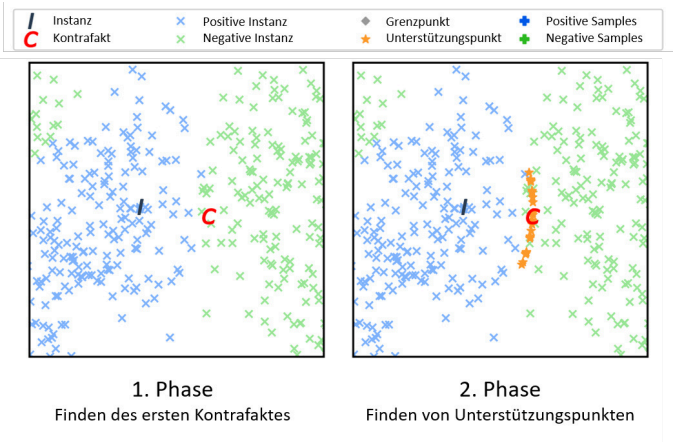


Abbildung 5.2: Phase 1 und Phase 2 des Vorgehens am Datensatz *Cancer* [Quelle: Eigene Darstellung]

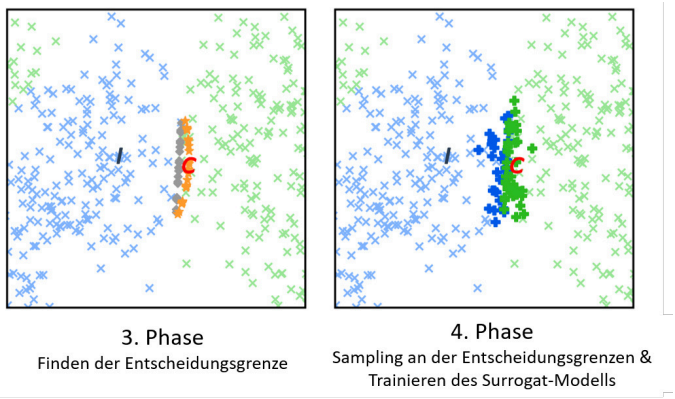


Abbildung 5.3: Phase 3 und Phase 4 des Vorgehens am Datensatz *Cancer* [Quelle: Eigene Darstellung]

5.3.5 Erklärbarkeit

Um exemplarisch die Erklärbarkeit der Modelle darzustellen, werden nachfolgend Beispiele aufgeführt und erläutert. Dabei wird zunächst der Anwendungsfall eines Erkennungssystems für Netzwerkangriffe betrachtet. Dieses System erkennt Angriffe auf ein Netzwerk und generiert dementsprechend Alarme für einen Anwender wie bspw. dem Systemadministrator. Für diese Alarme, insbesondere Fehlalarme, sollen Erklärungen generiert werden, sodass der Anwender diese nachvollziehen kann. Angriffserkennungssysteme erzeugen häufig Fehlalarme [Mei07], die durch entsprechende Erklärungen entsprechend zügiger als solche erkannt werden können. Nachfolgend werden drei unterschiedliche Arten von Erklärungen dargestellt, die bereits in Abschnitt 5.2.5 erläutert wurden.

Die Darstellung in Abb. 5.4 zeigt die Gewichte des linearen Modells. Dabei sind die Gewichte der einzelnen Merkmale erkenntlich. Bei den Merkmalen handelt es sich um spezielle Merkmale aus dem Bereich Datenraten und Netwerkverkehr.

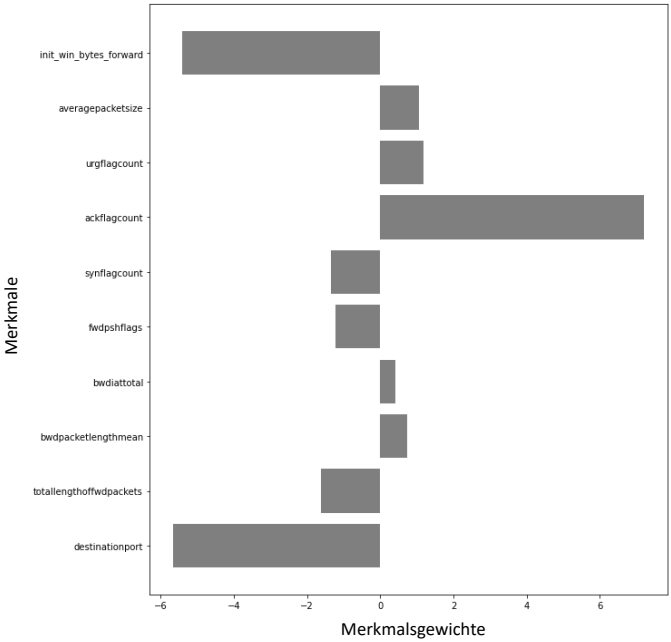


Abbildung 5.4: Relevante Merkmale der faktischen Instanz [Quelle: Eigene Darstellung].

Die relative Differenz in Abb. 5.5 zeigt die Ergebnisse der faktischen Instanz (schwarz) und der kontrafaktischen Instanz (rot). Hierbei wird ersichtlich, welche Änderungen bei den Merkmalen hätten durchgeführt werden müssen, um eine veränderte Vorhersage – oder in diesem Falle keinen Fehlalarm – zu erzeugen.

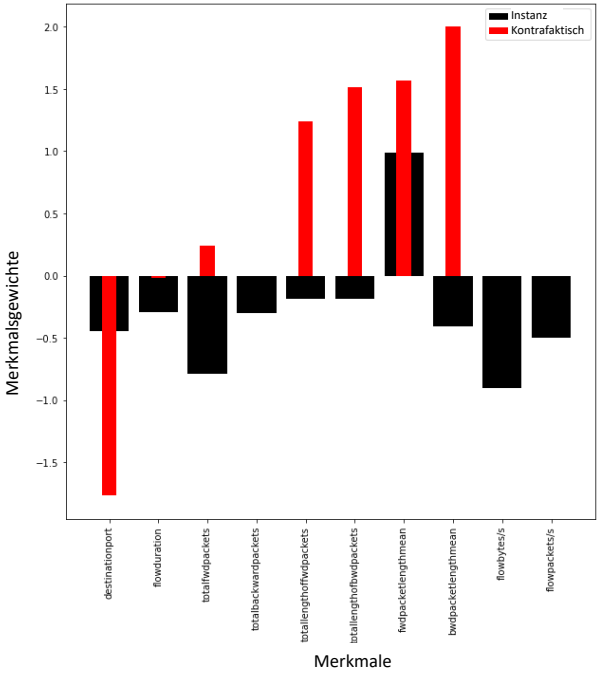


Abbildung 5.5: Relative Differenz zwischen kontrafaktischer und ursprünglicher Instanz [Quelle: Eigene Darstellung].

Die Darstellung der Erklärung, die das lokale Surrogat-Modell in Form eines Entscheidungsbaums liefert, unterstützt den Anwender dabei, den Entscheidungsprozess im Detail verstehen zu können. Für diese Aufgabe eignet sich der in Abb. 5.6 dargestellte Entscheidungsbaum, da der Entscheidungspfad der Vorhersage intuitiv verstanden werden kann [Elb19].

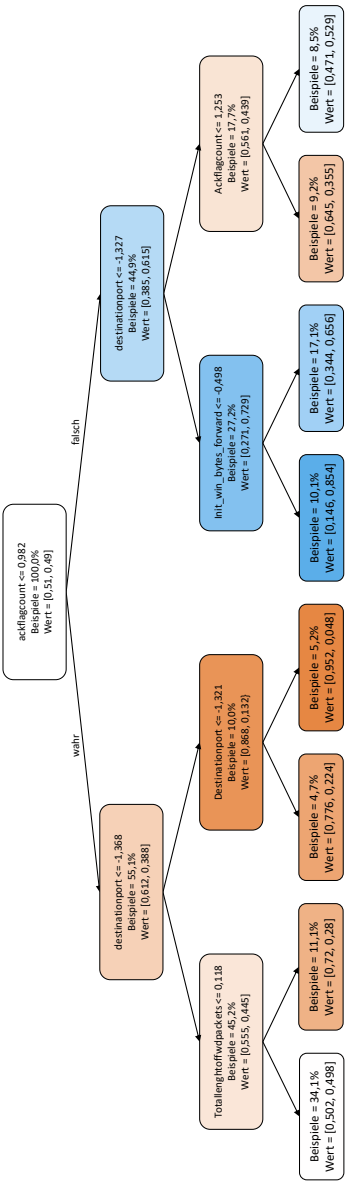


Abbildung 5.6: Lokales Surrogat-Modell als Entscheidungsbaum (Datensatz: KDD) [Quelle: In Anlehnung an [Bur21a], S. 89].

5.4 Zusammenfassung

In diesem Kapitel wurde ein phasenweises Vorgehen eines Verfahrens zur Erzeugung einer Erklärung, die in der Nähe einer Entscheidungsgrenze liegt, beschrieben. Das Vorgehen umfasst insgesamt fünf Phasen. Das Verfahren gibt die Möglichkeit unterschiedliche Arten von Erklärungen zu generieren, die den Anwender bei der Nachvollziehbarkeit der Ergebnisse unterstützen sollen. Zum einen zeigt eine Erklärung die Darstellung des lokalen Surrogates, das als lineares Modell trainiert wurde. Eine weitere Erklärung bildet die *relative Differenz* ab und zeigt die Ergebnisse der faktischen Instanz und der kontrafaktischen Instanz. Dabei wird ersichtlich, welche Änderungen bei den Merkmalen hätten durchgeführt werden müssen, um ein anderes Klassifikationsergebnis zu erzeugen. Die letzte Erklärung zeigt das lokale Surrogat-Modell als Entscheidungsbaum und unterstützt den Anwender dabei, den Entscheidungsprozess simulieren zu können. Das beschriebene Verfahren wurde im Rahmen einer Evaluation mit einem Referenzverfahren verglichen. Die Ergebnisse machten deutlich, dass das Verfahren – verglichen mit dem Referenzverfahren *LIME* – eine höhere Modelltreue generiert. Zudem zeigte sich, dass das lokale Surrogat, das als Entscheidungsbaum trainiert wurde, häufiger eine höhere Modelltreue hatte als das Surrogat, das als lineares Modell trainiert wurde.

Nachdem im Kapitel 4 und 5 Verfahren zur Erzeugung von Erklärungen entworfen wurde, wird im nachfolgenden Kapitel die Nachvollziehbarkeit von Erklärungen im Rahmen von Nutzerstudien untersucht.

6 Benutzerstudien am Beispiel unterschiedlicher Domänen und Anwendergruppen

Im Fokus dieses Kapitels stehen die durchgeführten Nutzerstudien. Nachdem in den vorangegangenen Kapiteln Verfahren entwickelt wurden, die die Nachvollziehbarkeit komplexer Modelle unterstützen, wird in diesem Kapitel untersucht, welche Art von Erklärungen die Nachvollziehbarkeit unterstützen, um dadurch Vertrauen in die Modelle aufbauen zu können. Dazu wurden drei unterschiedliche Benutzerstudien konzipiert und durchgeführt.

6.1 Problemstellung

Eine der großen Herausforderungen im Bereich der erklärbaren KI ist es, festzustellen, welche Eigenschaften Erklärungen enthalten müssen, um von Anwendergruppen verstanden zu werden. In Bezug auf ein Modell, das eine Vorhersage liefert, soll eine Erklärung die Gründe einer Entscheidung so darstellen, dass diese von einem Anwender nachvollzogen werden können. In der Regel wird Sachverhalten, die nicht verstanden werden, nicht vertraut. Das Verständnis und Vertrauen basierend auf vordefinierten Metriken zu messen, ist ein offenes Forschungsthema und bislang nicht gelöst. Daher sind Benutzerstudien, die die Nachvollziehbarkeit von Erklärungen untersuchen, ein essenzieller Baustein. Im Folgenden wird zunächst ein Überblick über die drei durchgeführten Benutzerstudien gegeben.

Für die ersten beiden Benutzerstudien wurde eine Versuchsaufgabe entworfen, die mit unterschiedlichen Erklärungen und Anwendergruppen untersucht wurde. Die Aufgabe war so aufgebaut, dass die Versuchsteilnehmer dazu aufgefordert wurden, eine Schätzung zu einer bestimmten Aufgabe abzugeben. Im Anschluss daran erhielten die Versuchsteilnehmer eine KI-Unterstützung in Form einer Vorhersage und ggfs. einer lokalen oder globalen Erklärung. Anschließend hatten die Versuchsteilnehmer die Möglichkeit, ihre erste Schätzung anzupassen, den Empfehlungen der KI-Unterstützung zu folgen und somit eine zweite Schätzung abzugeben.

Die erste Benutzerstudie befasste sich mit der Nachvollziehbarkeit lokaler und globaler Erklärungen, bei denen die Merkmalsrelevanz in unterschiedlicher Form dargestellt wurde. Der Versuchsteilnehmer ist ein Domänenexperte.

Die zweite Benutzerstudie konzentrierte sich ebenso auf die Nachvollziehbarkeit lokaler und globaler Erklärungen, die im Gegensatz zur ersten Benutzerstudie nur auf einer Form der Darstellung der Merkmalsrelevanz basieren. Die Nutzerstudie wurde mit insgesamt 80 Versuchsteilnehmern durchgeführt.

Die dritte Benutzerstudie untersuchte, wie explizite Merkmale, die bspw. durch den Einsatz von Ontologien erzeugt werden, beim Verständnis unterstützen können. Sowohl bei der ersten als auch bei der zweiten Benutzerstudie wurde festgestellt, dass das Verständnis der Merkmale über die Nachvollziehbarkeit einer Erklärung entscheidet. Dies bedeutet, dass eine Kombination von Merkmalen unter Umständen nachvollziehbarer ist, als diese einzeln darzustellen. Dieser Sachverhalt wurde im Rahmen der Benutzerstudie untersucht. Die nachstehende Abb. 6.1 gibt einen Überblick über die durchgeführten Nutzerstudien.

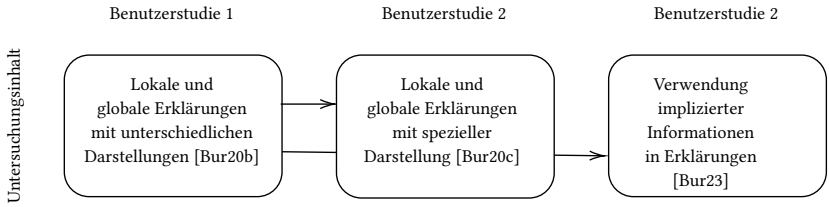


Abbildung 6.1: Überblick über die Untersuchungsinhalte der Benutzerstudien. Ergebnisse der Benutzerstudie 1 werden in den Benutzerstudien 2 und 3 näher untersucht [Quelle: Eigene Darstellung].

6.2 Nutzerstudie 1

Um verschiedene Arten von Erklärungen mit der Versuchsperson zu untersuchen, wurde eine spezielle Aufgabenstellung konzipiert, die in der maritimen Domäne relevant ist und bei der eine KI-Unterstützung durch eine Vorhersage basierend auf einem Blackbox-Modell sinnvoll ist. Nachfolgend werden zunächst die Datengrundlage und das verwendete Vorhersagemodell beschrieben. Das Vorhersagemodell basiert auf den Arbeiten von [Ann20] und wurde für die Erzeugung der Erklärungen verwendet.

6.2.1 Die Datengrundlage

AIS-Daten dienen u. a. dazu, Kollisionen zwischen Schiffen zu verhindern, den Küstenstaaten Informationen über die Schiffe und ihre Ladungen zu liefern und den illegalen Fischfang zu kontrollieren [Tet05]. Die Trainingsdaten enthalten ca. 2 Milliarden dynamische Datenpunkte und ca. 200.000 Schiffe. Für das Training des Netzes wurden spezielle Merkmale generiert. Darunter sind die *Schiffstrajektorie*, die *zeitliche Differenz*, in der ein Schiff AIS-Daten sendet, die *Entfernung zur Küste*, die *Entfernung zum Hafen*, die *Geschwindigkeit* und der *Kurs*. Die Schiffstrajektorie ergibt sich aus den Positionsdaten, die dem globalen Navigationssatellitensystem entnommen werden. Die Entfernung zur Küste approximiert die Entfernung zum nächstgelegenen Punkt

auf der Küstenlinie, während die Entfernung zum Hafen die Entfernung zum nächstgelegenen Hafen angibt. Die Schiffstypen wurden beim Training des Modells in folgende Klassen unterschieden: Frachttanker, Fischerboot, Passagierschiffe, Sportboote und Schlepperschiffe. Als zugrundeliegendes Vorhersagemodell wurde das residuale Netz basierend auf den Arbeiten von Anneken et al. [Ann20] verwendet.

6.2.2 Erklärungen

In diesem Abschnitt werden die Erklärungen, die innerhalb der Benutzerstudie verwendet wurden, dargestellt. Die zugrundeliegende algorithmische Vorgehensweise der Erklärungen wurde bereits in Abschnitt 2.5.3 erläutert. Nachfolgend wird auf die Ergebnisse der Verfahren eingegangen. Die nachstehende Tabelle 6.1 veranschaulicht die Instanz anhand der Merkmale, die zur Generierung der lokalen Erklärungen verwendet wurde.

Tabelle 6.1: Test-Instanz aus der Klasse *Frachttanker* [Quelle: In Anlehnung an [Bur20b], S. 8].

Merkmale	Merkmalswerte
Trajektorie X	47,66°
Trajektorie Y	−3,9°
zeitliche Differenz	80 Sekunden
Distanz zur Küste	20 Kilometer
Distanz zum Hafen	2450 Kilometer
Geschwindigkeit	3,8 Knoten
Kurs	1,58°N

Die Ausgabe der SHAP-Erklärung ist in Abb. 6.2 dargestellt. Es ist zu erkennen, dass die Klasse *Frachttanker* mit einer Wahrscheinlichkeit von 0,48 vorhergesagt wird. Der Basiswert ist die durchschnittliche Ausgabe über den

Trainingsdatensatz und beträgt 0,46. Den größten Einfluss auf die Vorhersage *Frachttanker* hat das Merkmal *Geschwindigkeit*. Andere Auswertungen für die Vorhersage der Klasse *Fischerboot* zeigte, dass die wichtigsten Merkmale für diese die *Entfernung zur Küste* und die *Entfernung zum Hafen* sind. Das bedeutet, dass die Merkmale *Entfernung zur Küste* und *Entfernung zum Hafen* bei der Klasse *Fischerboot* dominanter sind als bei den anderen Klassen. Dies könnte darauf zurückzuführen sein, dass die meisten Fischerboote bspw. näher an der Küste fahren als z. B. der Schiffstyp Frachttanker. Wie in Abb. 6.3 zu sehen ist, gibt das Verfahren eine Merkmalsrelevanz aus. Dabei ist zu erkennen, dass der Kurs und die Geschwindigkeit die zwei ausschlaggebenden Merkmale sind. Die Ausgabe von LIME in Abb. 6.4 zeigt den Beitrag der einzelnen Merkmale zur Vorhersage. In der Abbildung ist zu erkennen, dass das Merkmal *Geschwindigkeit* entgegen der zwei anderen Erklärungen nicht zur Vorhersage *Frachttanker* beiträgt. Schiffe der Klasse *Frachttanker* fahren üblicherweise mit ca. 15-20 Knoten pro Stunde.

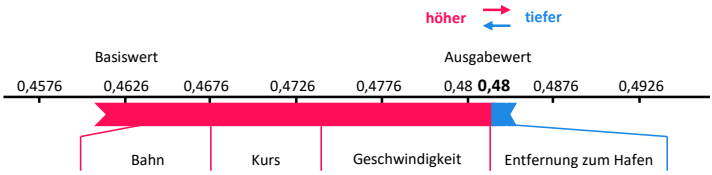


Abbildung 6.2: Erklärung *SHAP* für die Klasse *Frachttanker* [Quelle: In Anlehnung an [Bur20b], S. 5].

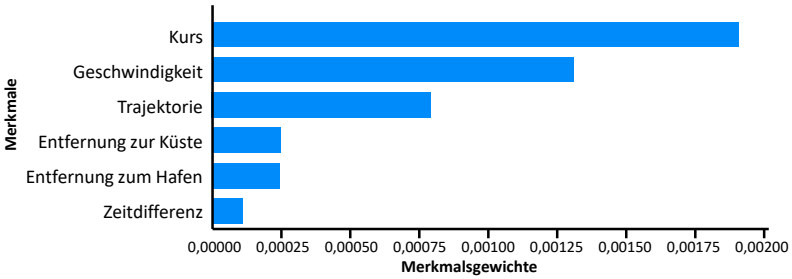


Abbildung 6.3: Erklärung *MCR* für *Frachttanker* [Quelle: In Anlehnung an [Bur20b], S. 6].

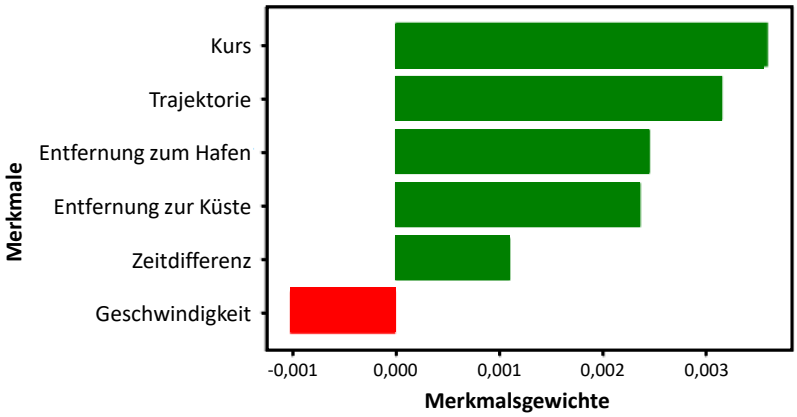


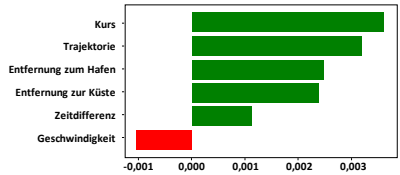
Abbildung 6.4: Erklärung *LIME* für Test-Instanz *Frachttanker* [Quelle: In Anlehnung an [Bur20b], S. 6].

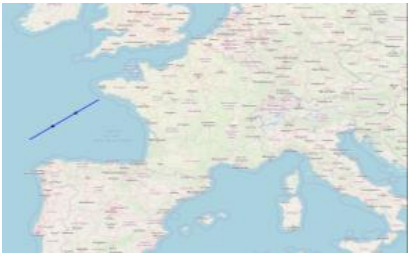
Die Erklärung sp-LIME in der nachstehenden Abb. 6.5 versucht, ein globales Verständnis des Modells zu erzeugen. In dem Beispiel werden drei repräsentative Instanzen und deren Erklärungen erzeugt. Zusätzlich wurden die Trajektorien der Beispielinstanzen hinzugefügt, um den Entscheidungsprozess

des Modells detaillierter abzubilden. Damit der Anwender einen Eindruck von dem Entscheidungsprozess erlangt, muss dieser die Testinstanz mit den generierten Erklärungen und Instanzen aus sp-LIME vergleichen. Für die Klasse *Frachttanker* waren die entscheidenden Merkmale *Kurs* und *Geschwindigkeit*.

Passagierschiff

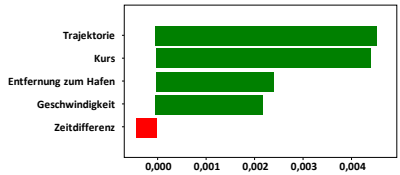
Zeit-differenz	Entfernung zur Küste	Entfernung zum Hafen	Geschwin-digkeit
250 Sek.	2,4 km	3939 km	4,7 Knoten





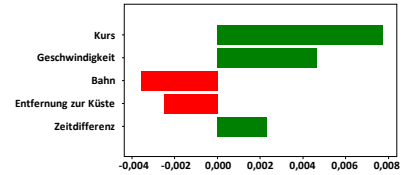
Fischerboot

Zeit-differenz	Entfernung zur Küste	Entfernung zum Hafen	Geschwin-digkeit
100 Sek.	17,6 km	912,5 km	0,44 Knoten



Frachtschiff

Zeit-differenz	Entfernung zur Küste	Entfernung zum Hafen	Geschwin-digkeit
120 Sek.	18,5 km	1353 km	5,8 Knoten



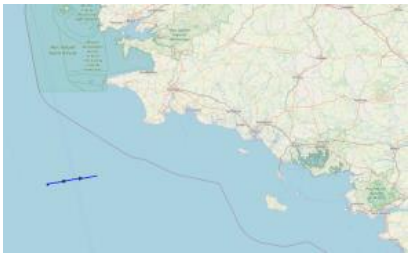


Abbildung 6.5: Erklärung sp-LIME mit den Klassen *Frachttanker*, *Passagierschiff* und *Fischerboot* [Quelle: In Anlehnung an [Bur20b], S. 7].

Für den lokalen Ansatz SHAP und die globalen Ansätze MCR und sp-LIME waren die Merkmale *Geschwindigkeit* und *Kurs*, die mit dem grössten Einfluss. Für den lokalen Ansatz LIME waren der *Kurs* und *Trajektorie* die essenziellen Merkmale, die zur Vorhersage beigetragen haben. Die *Geschwindigkeit* beeinflusste die Vorhersage der Klasse *Frachttanker* negativ.

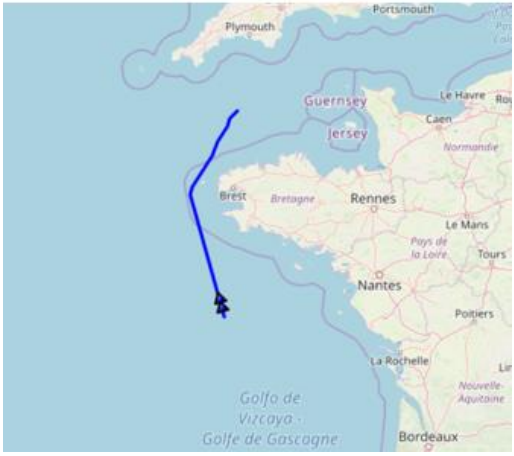
6.2.2.1 Versuchsaufbau

Bei dem Versuch wurde untersucht, ob der Versuchsteilnehmer seine erste Vorhersage anpasst, wenn dieser die Unterstützung von einem KI-System erhält. Diese wurde in Form der eigentlichen Vorhersage und einer Erklärung gegeben. Insgesamt wurden vier Versuchsausprägungen entworfen, bei denen zunächst immer der Versuchsteilnehmer seine Einschätzung abgeben musste. Jede der vier Versuchsausprägungen (A1, A2, A3, A4) enthielt eine andere Art der Erklärung. Dabei sollte ermittelt werden, welche Art der Erklärung den Versuchsteilnehmer unterstützt und ob dieser seine erste Vorhersage entsprechend der Empfehlung anpassen wird. Das Experiment wurde mittels eines Fragebogens durchgeführt. Dem Teilnehmer wurden grundlegende Fakten über das entsprechende Vorhersagemodell mitgeteilt, und es wurde ihm erklärt, wie das Modell zur Vorhersage des Schiffstyps gelangt. Darüber hinaus wurde erläutert, dass das verwendete Modell eine Blackbox ist und nicht nachvollzogen werden kann, wie genau die Vorhersage aus dem Modell abgeleitet wurde. Der Versuchsteilnehmer wurde in die Grundlagen des erklärbaren maschinellen Lernens eingeführt und darüber aufgeklärt, wie Erklärungen bei der Nachvollziehbarkeit von Vorhersagen einer Blackbox unterstützen können.

6.2.3 Versuchsaufbau

Für jede der insgesamt vier Versuchsausprägungen musste der Versuchsteilnehmer den Schiffstyp basierend auf der Schiffstrajektorie und weiteren Merkmalen, die der Abb. 6.6 entnommen werden können, vorhersagen. Dazu wurden vier Aufgaben erstellt. Die möglichen Schiffstypen, aus denen

der Versuchsteilnehmer wählen konnte, waren Frachttanker, Fischerboot, Passagierschiffe, Sportboote und Schlepperschiffe.



Zeitdifferenz	Entfernung zur Küste	Entfernung zum Hafen	Geschwindigkeit
80 Sek.	20 km	2450 km	5,8 Knoten

Abbildung 6.6: Beispiel der Aufgabe: Darstellung der Trajektorie und der weiteren Merkmale [Quelle: In Anlehnung an [Bur20b], S. 8].

Weiter wurde der Versuchsteilnehmer darum gebeten, diejenigen Merkmale anzugeben, die für seine Entscheidung am ausschlaggebendsten waren. Nach der ersten Schätzung erhielt der Versuchsteilnehmer eine Unterstützung, die einen der fünf Schiffstypen vorhersagte. Außerdem wurde ihm bei jeder Aufgabe eine der vier Erklärungen vorgelegt, die die Vorhersage nachvollziehbarer gestalten sollte. Die erste Aufgabe basierte auf der Erklärung SHAP (A1), die zweite Aufgabe auf der Erklärung MCR (A2), die dritte Aufgabe auf der Erklärung LIME (A3) und die vierte Aufgabe auf der Erklärung sp-LIME (A4).

Im Anschluss daran hatte der Teilnehmer die Möglichkeit, seine erste Schätzung auf der Grundlage der Unterstützung anzupassen. Außerdem sollte der Teilnehmer beurteilen, ob die gegebene Erklärung verständlich war und ihn bei der Lösung der Aufgabe unterstützt hat. Am Ende des Versuchs wurden allgemeine Fragen zur gesamten Thematik abgefragt.

6.2.4 Ergebnisse

Die nachstehende Tabelle 6.2 zeigt für jede Aufgabe die erste Schätzung des Versuchsteilnehmers, die Vorhersage der Blackbox und die zweite Schätzung. Die Grundwahrheit gibt das tatsächliche Ergebnis an, dieses wurde so ausgewählt, dass es stets mit der Vorhersage der Blackbox übereinstimmt. In Aufgabe 1 änderte der Versuchsteilnehmer seine zweite Schätzung und folgte nicht der Empfehlung der Unterstützung. In Aufgabe 2 wählte der Versuchsteilnehmer zwei Schiffstypen, die KI-Empfehlung bestätigte eine Auswahl des Teilnehmers, und dieser passte seine Entscheidung daraufhin an. In Aufgabe 3 zeigte sich, dass der Experte erneut zwei Schiffstypen wählte und die KI eine seiner Entscheidungen unterstützte. Der Experte behielt jedoch in der zweiten Schätzung ebenfalls beide Schiffstypen. In Aufgabe 4 passte der Teilnehmer seine Schätzung entsprechend der Empfehlung der KI an.

Tabelle 6.2: Ergebnisse pro Versuchsausprägung [Quelle: In Anlehnung an [Bur20b], S. 9].

	Schätzung 1	AI-Vorhersage	Schätzung 2
A1	Schlepperschiff	Frachttanker	Sportboot
A2	Fischerboot Sportboot	Fischerboot	Fischerboot
A3	Schlepperschiff Sportboot	Sportboot	Schlepperschiff Sportboot
A4	Frachttanker	Sportboot	Sportboot

Nachdem die Aufgaben abgeschlossen waren, wurden dem Versuchsteilnehmer die folgenden Fragen gestellt:

- Was gefiel Ihnen an der Darstellung der Erklärungen?
- Was gefiel Ihnen nicht an der Darstellung der Erklärungen?
- Welche Erklärung war besonders nachvollziehbar?
- Gibt es Verbesserungspotenziale? Wenn ja, welche?
- Waren die Erklärungen generell hilfreich? (Likert-Skala 1-7)
- Würden Sie einer KI-Empfehlung mit einer Erklärung mehr vertrauen als einer ohne? (Likert-Skala 1-7) [Bur20b]

An der Darstellung der Erklärung gefiel dem Teilnehmer besonders, dass die Relevanz der unterschiedlichen Merkmale dargestellt wurde. Die Bedeutung der Merkmale wurde vom Teilnehmer negativ bewertet. Dabei war nicht ausführlich beschrieben, ob das Merkmal *Entfernung zum Hafen* die Entfernung zum Hafen des Bestimmungsortes oder zum nächstgelegenen Hafen ist. Für eine Erklärung ist demnach eine genaue Beschreibung des Merkmals entscheidend. Sind die Merkmale für den Anwender nicht nachvollziehbar, kann auch eine Erklärung, die diese Merkmale enthält, nicht nachvollziehbar sein. Die favorisierte Erklärung des Versuchsteilnehmers war *sp-LIME*, da versucht wurde, ein Modellverständnis zu geben. Bei dieser Erklärung korrigierte der Teilnehmer auch seine erste Schätzung entsprechend der KI. Zur Verbesserung schlug der Teilnehmer vor, unterschiedliche Merkmale weiter zu kombinieren und somit eine implizite Information zu erzeugen. Der Teilnehmer empfahl bspw., die Trajektorie mit dem Kurs zu kombinieren. Bei den zwei letzten Fragen wurde eine siebenstufige Likert-Skala verwendet. Der Teilnehmer gab bei der Frage, wie hilfreich die Erklärungen generell sind, an, dass er diesbezüglich unentschlossen (Likert-Skala-Wert (4)) sei. Dies könnte u. U. auf die unklare Beschreibung der Merkmale zurückzuführen sein. Der Teilnehmer bekräftigte jedoch, dass er im Allgemeinen einer KI-Empfehlung mit Erklärung mehr vertrauen würde als einer ohne Erklärung (Likert-Skala-Wert (7)).

6.2.5 Zusammenfassung

In diesem Abschnitt wurde eine Benutzerstudie mit einem Experten, einem Offizier, innerhalb der maritimen Domäne durchgeführt. Die Aufgabe des Versuchsteilnehmers bestand darin, anhand einer Schiffstrajektorie und weiterer Merkmale (wie z. B. dem Kurs und der Geschwindigkeit) einen bestimmten Schiffstyp zu schätzen. Nachdem der Versuchsteilnehmer die Zuweisung durchgeführt hatte, wurde diesem eine Unterstützung durch ein KI-System gegeben. Diese zeigte dem Versuchsteilnehmer die Vorhersage des Schiffstyps und eine Erklärung an. Basierend darauf hatte der Versuchsteilnehmer die Möglichkeit, seine erste Schätzung des Schiffstyps anzupassen und somit eine zweite Schätzung abzugeben. Insgesamt wurden vier Aufgaben erstellt, wobei jede eine andere Art der Erklärung enthielt. Die Ergebnisse zeigten, dass der Versuchsteilnehmer in 50 % der Fälle seine erste Schätzung revidierte und den Empfehlungen der KI folgte. Die Merkmalsrelevanz wurde als hilfreich angesehen und die Erklärung durch *sp-LIME*, bei der der globale Entscheidungsprozess dargestellt wurde, wurde gegenüber den lokalen Erklärungen bevorzugt. Problematisch stellte sich die Definition der Merkmale für den Versuchsteilnehmer dar, da diese für ihn nicht klar formuliert waren. Es ist daher wichtig, bei der Erstellung neuer Merkmale darauf zu achten, dass diese möglichst eindeutig von den Anwendern verstanden werden können. Als Fazit der ersten Nutzerstudie lässt sich die Aussage treffen, dass es einer Anwendergruppe mit Fachwissen über eine Domäne wichtig ist, den gesamten Entscheidungsprozess einer Blackbox nachvollziehen zu können und nicht nur einzelne Entscheidungen.

Basierend auf den Ergebnissen der ersten Nutzerstudie wird in der zweiten Nutzerstudie der Aspekt lokaler und globaler Erklärungen anhand der präferierten Darstellung der Merkmalsrelevanz mit einer größeren Menge an Versuchsteilnehmern untersucht.

6.3 Nutzerstudie 2

In dieser Benutzerstudie wurde untersucht, wie Versuchsteilnehmer ihre erste Vorhersage bei einer bestimmten Aufgabe anpassen, wenn diese zusätzlich eine Unterstützung in Form einer Erklärung erhalten. Das Vorgehen ist identisch mit dem der Benutzerstudie 1, jedoch war die Aufgabenstellung anders formuliert. In der Benutzerstudie 2 sollten die Teilnehmer die Noten von Schülern schätzen (Regressionsproblem). Nachdem die Teilnehmer die erste Schätzung der Note abgegeben hatten, erhielten diese eine KI-Unterstützung, die die Vorhersage der Note und eine Erklärung darstellte. Damit sollte herausgefunden werden, wie die Benutzer ihre erste Vorhersage auf der Grundlage der Art der Erklärungen anpassen. Insgesamt wurden vier verschiedene Versuchsausprägungen konzipiert.

6.3.1 Die Datengrundlage

Als Trainingsdaten wurde der Datensatz der Universität von Minho [Cor08] gewählt, der Informationen zu Schülern einer Sekundarstufe von insgesamt zwei portugiesischen Schulen enthält. Der Datensatz umfasst 649 Instanzen, jede einzelne Instanz spiegelt einen Schüler wider und umfasst neben den Noten der Schüler ebenso demografische, soziale und weitere schulische Merkmale. Für das Training des Modells wurde eine Teilmenge von 14 Merkmalen aus den insgesamt 33 Merkmalen gewählt. Diese wurden entsprechend ihrer Nachvollziehbarkeit gewählt, um für die Versuchsteilnehmer möglichst leicht verständlich zu sein. Die Teilnehmer sollten nicht mit einer Vielzahl an Merkmalen überfordert werden. Dies wurde in einer Vorstudie untersucht und sichergestellt. Folgende Merkmale wurden verwendet, um das Vorhersagemodell zu trainieren: Geschlecht, Alter, Status der Eltern, Grund für die Wahl der Schule, Lernzeit (wöchentlich), Misserfolge, Aktivitäten, Beziehungsstatus, familiäre Situation, Freizeit (Stunden) pro Tag, Ausgehen (Stunden) pro Tag, Alkoholkonsum am Wochenende, Gesundheitsstatus, Abwesenheit vom Unterricht (Stunden).

6.3.2 Vorhersagemodell und Erklärungen

Insgesamt wurden zwei Vorhersagemodelle trainiert, um die Noten der Schüler vorherzusagen. Somit liegt das Problem der Regression vor, da ein kontinuierlicher Wert vorhergesagt wird (s. Abschnitt 2.1.2.2). Zum einen wurde ein lineares Modell als Whitebox-Modell und zum anderen ein Random Forest trainiert, der die Blackbox darstellt. Beide Verfahren wurden bereits in Abschnitt 2.1.2 beschrieben. Der mittlere absolute Fehler des linearen Modells beträgt 4,64 und der des Random-Forest-Modells 2,90. Somit sind die Vorhersagen des Blackbox-Modells genauer als die der Whitebox. Dies wurde gezielt so gewählt.

Insgesamt wurden vier Versuchsausprägungen erstellt. Jede enthielt eine andere Art der Erklärung. Darunter fallen drei Typen in die Kategorie eines Erklärtyps entsprechend Kapitel 3. Die zugrundeliegende algorithmische Vorgehensweise der Erklärverfahren wurde bereits in Abschnitt 2.5 erläutert. Als weiterer Typ einer Unterstützung wurde die Vorhersage der Blackbox ohne eine Erklärung eingebunden. Im weiteren Verlauf werden die vier Versuchsausprägungen näher erläutert.

Die nachstehende Abb. 6.7 zeigt auf der rechten Seite die erste Versuchsausprägung in Form einer Merkmalsrelevanz. Die Unterstützung basiert auf einem linearen Regressionsmodell (E1) als Whitebox. Den Teilnehmern wurde neben der Vorhersage des Modells als Ergebnis der linearen Regression mitgeteilt, dass die folgenden Merkmale einen signifikanten (positiv: grün oder negativ: rot) Einfluss auf die Noten der Schüler haben. Die anderen Merkmale haben – je nach Modell – keinen signifikanten Einfluss (grau).

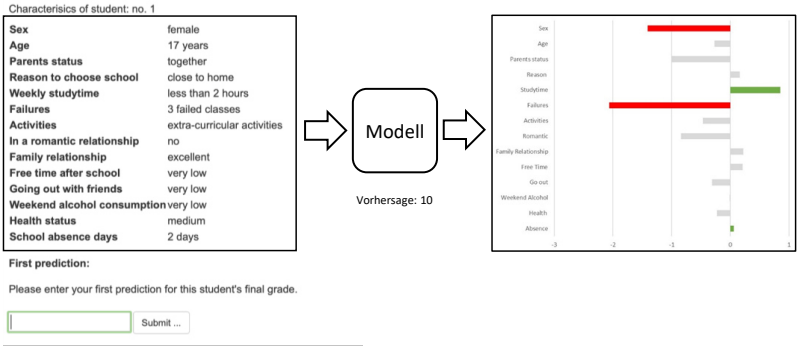


Abbildung 6.7: E1: Whitebox mit Vorhersage und Erklärung [Quelle: In Anlehnung an [Bur20c], S. 5].

In der folgenden Abb. 6.8 ist die zweite Versuchsausprägung zu sehen. Die Unterstützung basiert auf der Vorhersage eines Random Forest (E2) als Blackbox. Die Teilnehmer erhielten nur die Vorhersage (Note) für einen bestimmten Schüler ohne eine weitere Erklärung [Bur20c].

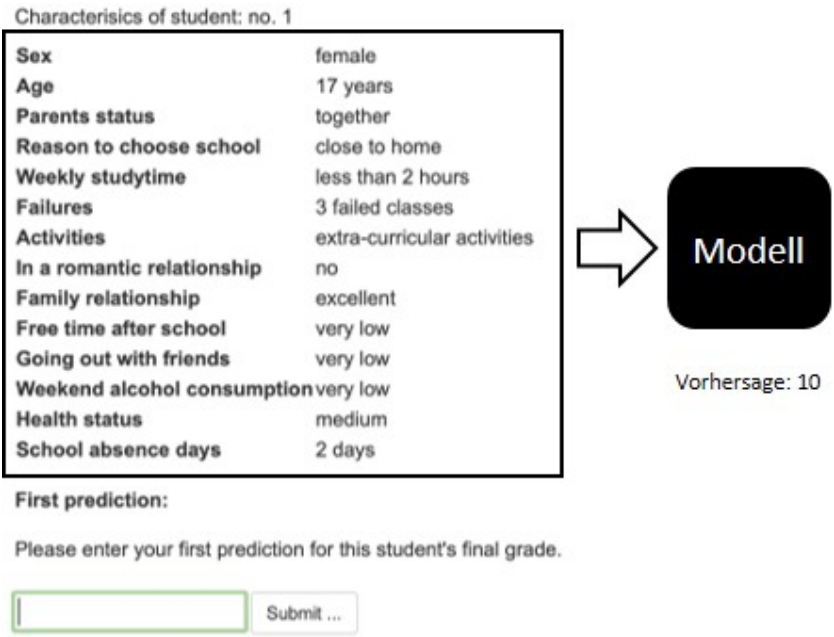


Abbildung 6.8: E2: Blackbox nur mit Vorhersage [Quelle: In Anlehnung an [Bur20c], S. 5].

Die nachstehende Abb. 6.9 stellt die dritte Versuchsausprägung dar. Die Unterstützung basiert auf einem Random Forest als Blackbox mit einer lokalen Erklärung durch LIME (E3). Den Teilnehmern wurde erklärt, dass für einen Einblick in den Entscheidungsfindungsprozess des Blackbox-Modells ein zusätzliches interpretierbares Surrogat-Modell generiert wurde. Dabei werden wie auch zuvor beim interpretierbaren Modell diejenigen Merkmale in unterschiedlichen Farben hervorgehoben, sofern diese für die Vorhersage ausschlaggebend bzw. nicht ausschlaggebend waren (positiv: grün oder negativ: rot).

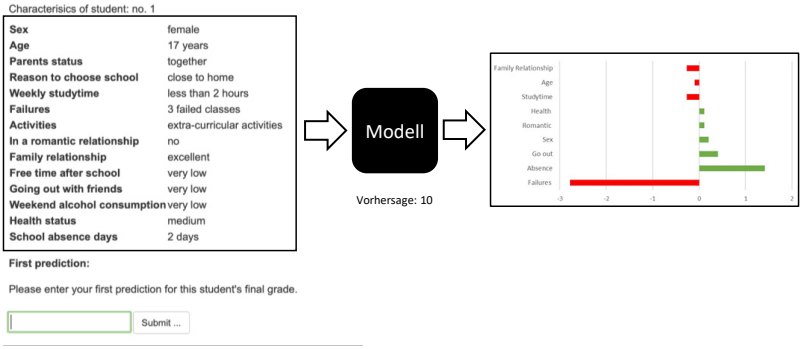


Abbildung 6.9: E3: Blackbox mit Vorhersage und lokaler Erklärung [Quelle: In Anlehnung an [Bur20c], S. 6].

Die Abb. 6.10 zeigt die vierte Versuchsausprägung. Die Unterstützung basiert auf einem Random Forest als Blackbox mit einer globalen Erklärung durch sp-LIME (E4). Den Teilnehmern wurde erklärt, dass drei repräsentative Schüler ausgewählt wurden, die bestmöglich den Entscheidungsprozess des Modells abzubilden versuchen.

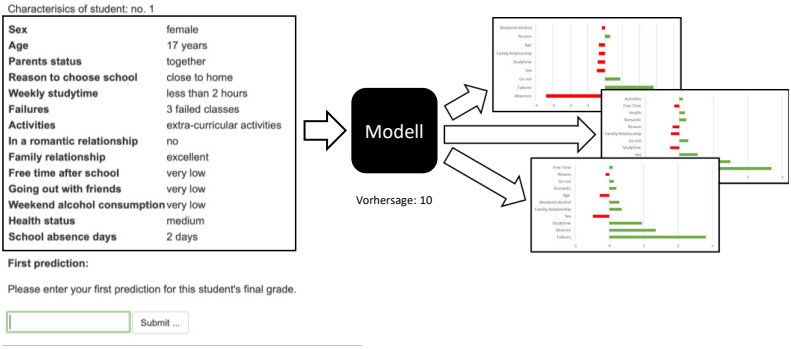


Abbildung 6.10: E4: Blackbox mit Vorhersage und globaler Erklärung [Quelle: In Anlehnung an [Bur20c], S. 6].

6.3.3 Versuchsaufbau

Den Teilnehmern jeder Versuchsausprägung wurden grundlegende Informationen über die Vorgehensweise des jeweiligen Verfahrens mitgeteilt sowie darüber, wie jedes Modell zur Erzeugung einer Vorhersage der Schülernote verwendet wird. Darüber hinaus wurde den Teilnehmern in den Versuchsausprägungen, bei denen ein Blackbox-Modell (E2-E4) verwendet wurde, erklärt, dass Ergebnisse eines Random Forest in der Regel nicht nachvollziehbar sind und daher als Blackbox bezeichnet werden.

In den Versuchsausprägungen, bei denen Verfahren zur Erzeugung einer Erklärung eingebunden wurden (E3, E4), wurden die Teilnehmer in die Grundlagen von Surrogat-Modellen eingeführt. Ebenso erfuhren sie, wie diese bei der Nachvollziehbarkeit einer Vorhersage unterstützen können. Anhand eines Beispiels wurde den Teilnehmern gezeigt, dass ein Surrogat-Modell verwendet werden kann, um die Vorhersage einer Blackbox (entweder global oder lokal) nachzuvollziehen. Weiter wurde herausgestellt, dass für die Entscheidung

essenzielle Merkmale (wie z. B. die akademischen und persönlichen Merkmale eines Schülers) hervorgehoben werden. Für die Versuchsausprägung der Whitebox (E1) wurde angegeben, dass das Modell einen mittleren absoluten Fehler (MAE) von 4,64 aufweist. Genauer wurde beschrieben, dass die durchschnittliche Abweichung des Modells bei 4,64 Notenpunkten von der wahren Note eines bestimmten Schülers abweicht. Die Blackbox in den Versuchsausprägungen (E2-E4) hat einen mittleren absoluten Fehler (MAE) von 2,9. Daher wurde erklärt, dass die Blackbox eine bessere Vorhersagegenauigkeit bei der Vorhersage der Note aufweist als die Whitebox, da die durchschnittliche Abweichung des Modells bei 2,9 Notenpunkten liegt. Dies wurde damit begründet, dass die Blackbox besser in der Lage ist, komplexere Sachverhalte zu erlernen, als die Whitebox.

Das Experiment wurde auf der Plattform Amazon Mechanical Turk[22] unter Verwendung von Sophie Labs[23] für die Gestaltung von Online-Experimenten durchgeführt. Insgesamt wurden 80 Teilnehmer für die Benutzerstudie rekrutiert. Diese wurden zufällig jeweils einer der vier Versuchsausprägungen zugeordnet. Somit konnte jeder Teilnehmer nur an einer Versuchsausprägung teilnehmen, was durch das System mit der Zuweisung einer eindeutigen Identifikationsnummer sichergestellt wurde. Vor Beginn des Experiments erhielten die Teilnehmer eine kurze Einführung in das Szenario und die Aufgabe. Den Versuchsteilnehmern wurde mitgeteilt, dass die Bezahlung daran angepasst wird, wie gut ihre Schätzung über die Noten der Schüler ist (d. h. die durchschnittliche absolute Abweichung von der tatsächlichen Note). Die Teilnehmer erhielten eine Teilnahmegebühr von 0,50 Euro. Im Versuch konnten sie mit ihren Schätzungen Punkte sammeln. Dabei entsprach ein Punkt = 0.01 Euro. Die maximale Anzahl von Punkten, die ein Teilnehmer mit einer Schätzung verdienen konnte, betrug 200. Das Vorgehen wurde deshalb gewählt, um sicherzustellen, dass die Teilnehmer einen Anreiz haben, eine möglichst gute Leistung zu erbringen.

[22] https://www.mturk.com/, letzter Abruf: 17.04.2021.
[23] https://www.sophielabs.com/, letzter Abruf: 03.05.2021.

6.3.3.1 Die Aufgabe

Die Aufgabe der Teilnehmer bestand darin, die Abschlussnote verschiedener Schüler vorherzusagen. Die Noten erstreckten sich über eine ganze Zahl zwischen 0 (schlechteste Note) und 20 Punkten (beste Note). Um eine Schätzung über die Note vornehmen zu können, wurden den Teilnehmern Merkmale der akademischen und persönlichen Eigenschaften der Schüler gegeben. Nachdem die Teilnehmer ihre Schätzung abgegeben hatten, erhielten sie Unterstützung in Form von einer der Versuchsausprägungen, die unter Abschnitt 6.3.2 vorgestellt wurden. Nach der Unterstützung erhielten die Teilnehmer die Möglichkeit, ihre erste Schätzung entsprechend anzupassen. Nach der Abfrage dieser zweiten Schätzung war die Aufgabe beendet. Insgesamt gab es fünf Aufgaben, in denen die Teilnehmer die Noten der Schüler schätzten und im Zuge dessen entsprechend anpassen konnten. Die nachstehende Abb. 6.11 ein Beispiel der Aufgabe, wie diese den Teilnehmer präsentiert wurde.

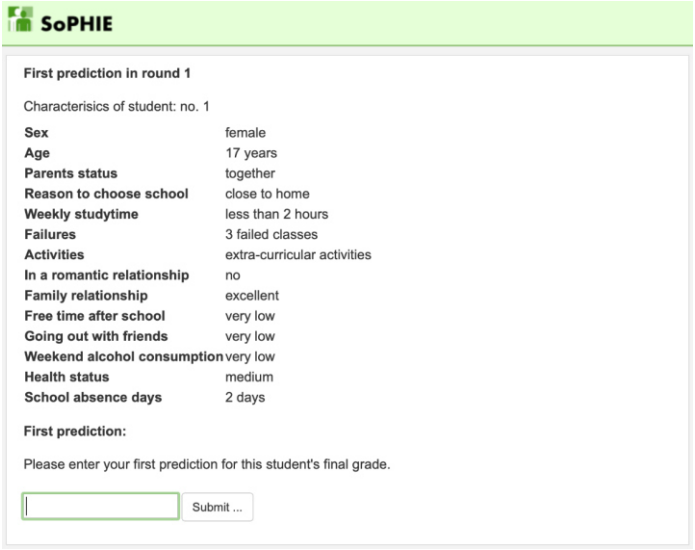


Abbildung 6.11: Beispiel einer Aufgabe: Darstellung der Merkmale eines Schülers und der Abfrage einer Schätzung [Quelle: In Anlehnung an [Bur20c], S. 11].

6.3.4 Die Metriken

Für die Analyse lagen insgesamt 400 Beobachtungen der 80 Teilnehmer über 5 Runden (20 Teilnehmer pro Versuchsausprägung) vor. Die Schätzgenauigkeit wurde an den Vorhersagefehlern gemessen, die als die absolute Abweichung der jeweiligen Vorhersage des tatsächlichen Wertes (Grundwahrheit) definiert

wurde. Der absolute Vorhersagefehler wurde mit absErr(μ) für die erste Vorhersage μ und absErr($\hat{\mu}$) für die zweite Vorhersage $\hat{\mu}$ angegeben. Dementsprechend wird der relative Vorhersagefehler mit

$$\text{relErr} = \frac{\text{absErr}(\hat{\mu})}{\text{absErr}(\mu)} \tag{6.1}$$

berechnet. Um das Vertrauen in die jeweilige Versuchsausprägung quantifizieren zu können, wurde die Differenz zwischen der anfänglichen Schätzung der Teilnehmer und der jeweiligen Anpassung durch eine entsprechende Unterstützung berechnet. Es wurde angenommen, dass eine höhere Anpassung mit mehr Vertrauen in die jeweilige Unterstützung verbunden ist. Basierend auf der ersten Schätzung μ_{ti} erfasst γ_{ti} die Anpassung der zweiten Schätzung μ'_{ti} an die gegebene Unterstützung $\hat{\mu}_{ti}$:

$$\gamma_{ti} = \frac{\mu'_{ti} - \mu_{ti}}{\hat{\mu}_{ti} - \mu_{ti}}, \tag{6.2}$$

wobei i der Teilnehmer und t die Vorhersage-Runde ist. γ_{ti} meint das Gewicht der Unterstützung, d. h. das Gewicht, das der Teilnehmer der Unterstützung, die er erhält, beimisst, und $1 - \gamma_{ti}$ bezeichnet dementsprechend das Gewicht, das dieser seiner ersten Schätzung beimisst.

6.3.5 Ergebnisse

Das durchschnittliche Alter der Teilnehmer betrug 37 Jahre. Insgesamt wurden 48 Männer, 31 Frauen und eine diverse Person befragt.

In der folgenden Tabelle 6.3 sind die Ergebnisse des Vorhersagefehlers und der Anpassung aufgeführt. Anhand dieser wird erkenntlich, dass der absolute Vorhersagefehler absErr(μ) für die erste Vorhersage aller vier Versuchsausprägungen auf einem vergleichbaren Niveau liegen, leicht erhöht für die Versuchsausprägung mit der Whitebox (siehe Tabelle 6.3 mit 6,48, 6,32, 6,01 bzw.

6,12). Alle Unterschiede sind statistisch nicht signifikant (alle p-Werte 0,33 oder höher). Darüber hinaus wurde eine geringere Anpassung in Versuchsausprägung E2 im Vergleich zu E1, E3 und E4 (E2 ist signifikant niedriger als E1, E3 und E4, alle p-Werte < 0,05) durchgeführt.

Tabelle 6.3: Absolute und relative Schätzfehler. Absolute Fehler geben die durchschnittliche absolute Abweichung vom wahren Wert bei der ersten (est1) und der zweiten Schätzung (est2) an. Die Differenz beschreibt die mittlere absolute Änderung zwischen dem ersten und zweiten Fehler, z. B. bedeutet ein positiver Wert eine bessere zweite Vorhersage. Die relative Anpassung gibt an, inwieweit die Unterstützung für die zweite Schätzung verwendet wurde (d. h. 0 bedeutet überhaupt nicht und 1 bedeutet sehr stark) [Quelle: In Anlehnung an [Bur20c], S. 7].

	E1	*E2*	*E3*	*E4*
Schätzung 1 absErr(μ)	6,48	6,32	6,01	6,12
Schätzung 2 absErr($\hat{\mu}$)	5,12	4,51	3,70	3,93
Differenz	1,36	1,81	2,31	2,19
relErr($\mu/\hat{\mu}$)	0,79	0,71	0,61	0,64
relative Anpassung	0,59	0,42	0,56	0,55

Die absoluten Fehler in der zweiten Vorhersage absErr($\hat{\mu}$) sind bei der Versuchsausprägung E1 am höchsten, bei E2 sind sie etwas niedriger und bei den Ausprägungen E3 und E4 sind sie auf einem vergleichbaren Niveau (5,12; 4,51; 3,7 bzw. 3,93). Der Vorhersagefehler bei der Ausprägung E1 ist signifikant höher als der Vorhersagefehler aller Ausprägungen mit einem Blackbox-Modell (E2, E3 und E4; p-Werte mit 0,041, 0.000 bzw. 0.000 einseitig). Darüber hinaus ist der Vorhersagefehler der Ausprägung E2 signifikant höher als der Vorhersagefehler der Ausprägung E3 (mit p-Werten 0,038 einseitig), aber nicht höher als der Vorhersagefehler der Ausprägung E4 (mit p-Werten 0,091 einseitig). Alle anderen Unterschiede sind statistisch nicht signifikant (alle p-Werte 0,36 oder höher).

Insgesamt verbesserten die Teilnehmer ihre Schätzungen, nachdem sie eine KI-Unterstützung erhielten – unabhängig von der Versuchsausprägung.

Die Unterschiede zwischen den absoluten Fehlern der ersten und der zweiten Schätzung sind signifikant mit p-Werten unter 0,001.

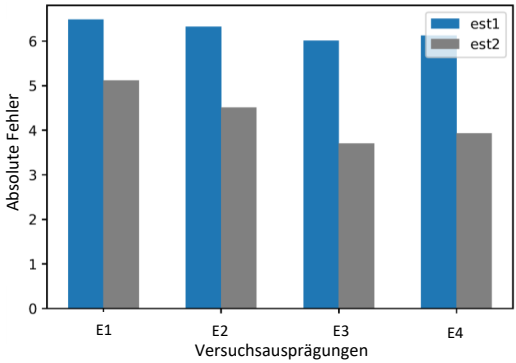


Abbildung 6.12: Absolute Fehler pro Versuchsausprägung (E1, E2, E3, E4), est1 = erste Schätzung und est2 = zweite Schätzung [Quelle: In Anlehnung an [Bur20c], S. 8].

Zusätzlich wurde *Cohens d* als Effektmaß für den Vergleich zwischen zwei Mittelwerten der relativen Anpassungen relErr($\mu/\hat{\mu}$) verwendet. Dabei wurden mittlere Effekte für den Vergleich von E1 vs. E2 (d=0,406) und kleine Effekte für E2 vs. E3 (-0,365) und E2 vs. E4 (-0,281) festgestellt, während nur marginale Effekte für E1 vs. E3 (0,074), E1 vs. E4 (0,083) und E3 vs. E4 (0,022) gefunden wurden.

Die Teilnehmer wurden aufgefordert, anhand einer Likert-Skala eine Rangfolge über die Nützlichkeit der gegebenen Erklärung zu erstellen. Die Skala reichte von 1 (trifft nicht zu) bis 7 (trifft voll zu). Die Box-Plots für E1, E2, E3, E4 sind in Abb. 6.13 pro Ausprägung mit den Mittelwerten 5,35, 5,45, 5,55 bzw. 5,43 dargestellt und zeigen keinen signifikanten Unterschied. Obwohl es nur geringe Unterschiede zwischen den Behandlungen gibt, wurde E3 als am nützlichsten bewertet.

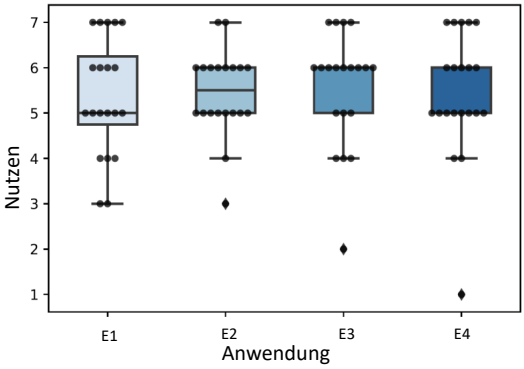


Abbildung 6.13: Box-Plots für Versuchsausprägung E1(5.35), E2(5.45), E3(5.55), E4(5.43) [Quelle: In Anlehnung an [Bur20c], S. 7].

6.3.6 Zusammenfassung

In diesem Abschnitt wurde eine Benutzerstudie mit insgesamt 80 Versuchsteilnehmern durchgeführt. Die Aufgabe bestand darin, die Abschlussnote von Schülern anhand unterschiedlicher Informationen eines Schülers zu schätzen. Nachdem die Versuchsteilnehmer die Schätzung abgegeben hatten, wurde ihnen eine Unterstützung durch ein KI-basiertes System gegeben. Die KI-Unterstützung zeigte jedem Versuchsteilnehmer die Vorhersage der Abschlussnote und ggfs. – je nach Versuchsausprägung – eine Erklärung an.

Durch den Versuch konnte festgestellt werden, dass alle Teilnehmer ihre Schätzung verbesserten, nachdem diese die KI-Unterstützung erhielten. Dies erfolgte unabhängig von der Versuchsausprägung. Es konnte beobachtet werden, dass das Whitebox-Modell nicht stärker in den Entscheidungsprozess einbezogen wurde als das Blackbox-Modell und dessen Erklärungen. Die Versuchsausprägung ohne eine Erklärung wies einen etwas geringeren absoluten Fehler verglichen mit dem Whitebox-Modell auf. Die Teilnehmer der Versuchsausprägung Blackbox ohne Erklärung passten ihre Schätzung

im Vergleich zur Blackbox mit einer lokalen Erklärung weniger an. Ein weiteres Ergebnis war, dass die Teilnehmer mit der Versuchsausprägung lokale Erklärungen ihre Schätzung mehr anpassten als die der globalen Erklärungen.

Wie Nutzerstudie 1 zeigte, ist das Verständnis der einzelnen Merkmal essenziell um eine gesamte Erklärung zu verstehen. Daher wird nachfolgend in der dritten Nutzerstudie der Aspekt der Kombination von Merkmalen in Form von expliziten Merkmalen um Erklärungen zu vereinfachen näher betrachtet.

6.4 Nutzerstudie 3

In der dritten Benutzerstudie wurde gezielt der Einfluss von Merkmalen in Erklärungen untersucht. Sowohl bei der ersten als auch bei der zweiten Benutzerstudie war es für die Versuchsteilnehmer essenziell, die zugrundeliegenden Merkmale, wie z. B. die Merkmale eines Schülers oder eines Schiffstyps, im Detail zu verstehen. Das Ziel dieser Studie war es, zu ergründen, welche Art von Merkmalen in einer Erklärung bevorzugt werden. Hierfür wurden zwei Arten von Merkmalen unterschieden: einfache und explizite Merkmale. Einfache Merkmale entstammen den ursprünglichen Trainingsdaten und werden ohne weitere Verarbeitungsschritte verwendet. Explizite Merkmale können bspw. durch die Verwendung von Ontologien (s. Abschnitt 2.3.4) aus den ursprünglichen Trainingsdaten zu neuen , expliziten Merkmalen, führen. Explizite Merkmale können, sofern das entsprechende Wissen bspw. in Form einer Ontologie vorliegt, in jegliche Art von Erklärung aus Kapitel 3 eingebunden werden. Im Gegensatz zu der ersten und zweiten Benutzerstudie wurde kein Vorhersagemodell und somit auch kein Trainingsdatensatz verwendet. Die Aufgaben wurden fiktiv erstellt.

6.4.1 Versuchsaufbau

Jede Aufgabe bestand aus den folgenden Informationen: eine Vorhersage über ein bestimmtes Ereignis und zwei unterschiedliche Arten von Erklärungen.

Insgesamt wurden fünf Aufgaben erstellt, die den Teilnehmern in einer zufälligen Reihenfolge gezeigt wurden. Am Ende der Studie wurden die Teilnehmer gebeten, die Gründe für ihre Entscheidung zu erläutern. In der Studie wurde als Maß ausschließlich die Präferenz der Teilnehmer bzgl. der Art der Erklärung gemessen. Den Teilnehmern wurden eine Vorhersage und zwei Erklärungen präsentiert, die sich hinsichtlich eines Aspektes unterschieden. Die Teilnehmer mussten sodann entscheiden, welche der beiden Erklärungen verständlicher für sie ist. Beide Erklärungen zeigten mindestens vier Merkmale. Diese unterschieden sich zwischen den beiden Erklärungen, wobei entweder einfache Merkmale, die nicht verändert wurden, oder ein explizites Merkmal verwendet wurde. Abhängig von der dargestellten Information kann dies die Art der präferierten Merkmale beeinflussen. Es wurden zwei Aspekte verglichen, die sich auf die Art der bevorzugten Merkmale in einer Erklärung auswirken können: der Schwierigkeitsgrad der Situation und die Komplexität der expliziten Merkmale.

Schwierigkeitsgrad der Situation: In manchen Situationen kann es entscheidend sein, alle detaillierten Informationen zu kennen, während in anderen ein grobes Verständnis ausreichend ist. Beispielsweise kann es in der medizinischen Domäne zwingend notwendig sein, über sämtliche Informationen zu verfügen, wohingegen im meteorologischen Umfeld nicht alle Informationen bekannt sein müssen.

Komplexität der expliziten Merkmale: Die Angabe bestimmter Merkmale kann in vielen verschiedenen Situationen zu einem umfassenderen Verständnis führen. Diese können aber in ihrer Komplexität unterschiedlich sein. Während es für eine Person einfach sein kann, schlusszufolgern, dass ein Datum im Juli der Jahreszeit Sommer zuzuweisen ist, kann es komplexer sein, ohne Vorwissen die Verbindung zwischen einem Breitengrad und einer Klimazone herzustellen. Dies ist subjektiv zu betrachten, da diese Beurteilung vom Vorwissen der jeweiligen Person abhängig ist.

6.4.1.1 Versuchsdesign

Für die Durchführung der Studie wurden Anwendungsbeispiele aus dem medizinischen und dem meteorologischen Bereich formuliert. Die Implikation, dass ein Datum im Juli der Jahreszeit Sommer zuzuweisen ist, wird bspw. als leichter angesehen als die Implikation, vom Breitengrad eines Ortes auf die entsprechende Klimazone zu schließen. Auch die Berechnung des Body-Mass-Index anhand des Gewichts und der Größe einer Person wird als schwieriger eingestuft, da hierfür die Formel bekannt sein muss. Den Teilnehmern wurde eine Ontologie zur Verfügung gestellt, die die Erzeugung der expliziten Merkmale beschreibt. Weitere Hilfsmittel waren nicht zulässig. Insgesamt wurden fünf Aufgaben (A-E) definiert wie in Tabelle 6.4 dargestellt, in denen die Teilnehmer die von ihnen präferierte Erklärung wählen sollten. Ein Beispiel für eine medizinische Situation gibt Listing 6.1.

Tabelle 6.4: Aus den dargestellten fünf Aufgaben sollten die Teilnehmer die bevorzugte Art der Ausprägung auswählen [Quelle: In Anlehnung an [Bur23]].

	Domäne	Situation	Merkmale	Unterschied
A	Med.	hoch	hoch	Wert BMI oder Kateg. BMI
B	Med.	hoch	hoch	Gewicht/Größe oder Kateg. BMI
C	Med.	hoch	hoch	Gewicht/Größe oder Wert BMI
D	Met.	niedrig	niedrig	Datum oder Jahreszeit
E	Met.	niedrig	hoch	Breitengrad oder Klimazone

```
Patient X hat kein Diabetes, weil...
Erklärung 1: der Glukose-test unter 123,5 mg/dL ausgefallen ist
    , der BMI des Patienten unter 25,0 liegt, er über 30 Jahre
    alt ist und einen Blutdruck unter 92 mm Hg besitzt.
Erklärung 2: der Glukose-test unter 123,5 mg/dL ausgefallen ist
    , der Patient ist nicht übergewichtig, er über 30 Jahre alt
     ist und einen Blutdruck unter 92 mm Hg besitzt.
```

Listing 6.1: Beispiel einer medizinischen Situation [Quelle: In Anlehnung an [Bur23]].

6.4.2 Ergebnisse

In diesem Abschnitt werden die Ergebnisse der Benutzerstudie präsentiert. Anschließend wird auf die Kommentare der Teilnehmer eingegangen, um ein besseres Verständnis über mögliche Präferenzen zu erhalten. Insgesamt wurden drei Hypothesen darüber getroffen, welche Art von Merkmalen in einer Erklärung bevorzugt werden:

Hypothese 1: Explizite Merkmale sind nachvollziehbarer.

Die meisten Menschen bevorzugen eine einfache Erklärung, die problemlos zu verstehen ist. Daher sollen explizite Merkmale versuchen, den kognitiven Aufwand für den Anwender zu reduzieren.

Hypothese 2: Detaillierte Merkmale werden in kritischen Situationen bevorzugt. Die Schwere einer Situation kann über den gewünschten Detailgrad von Merkmalen entscheiden. In kritischen Situationen, wie z. B. bei einem medizinischen Befund, waren die detaillierten Informationen mehr von Interesse als die zusammengefassten expliziten Merkmale.

Hypothese 3: Detaillierte Merkmale werden bevorzugt, wenn die expliziten Merkmale leicht zu erschließen sind. Sofern ein Mensch mit der Information vertraut ist und es ihm leicht fällt, die Implikationen selber abzuleiten, werden die detaillierten Merkmale präferiert.

Um diese Hypothesen zu überprüfen, werden in den folgenden Abschnitten die Ergebnisse der Benutzerstudie dargestellt und die Präferenzen zwischen den Merkmalen verglichen. Die zweite Hypothese wird überprüft, indem die meteorologische Aufgabe der medizinischen Aufgabe gegenübergestellt wird. Die dritte Hypothese wird analysiert, indem beide meteorologischen Themen miteinander verglichen werden und die Auswirkung des Schwierigkeitsgrades betrachtet wird. An der Benutzerstudie nahmen insgesamt 130 Teilnehmer teil, von denen 76,9 % männlich, 22,3 % weiblich und 0,8 % divers waren. Das Alter der Teilnehmer reichte von 18 bis 29 Jahren mit einem durchschnittlichen Alter von 22 Jahren. Die nachstehende Tabelle 6.5 zeigt die Ergebnisse der bevorzugten Erklärungen. In der Tabelle 6.5 sind die Ergebnisse der Abstimmung bzgl. der Präferenz im Hinblick auf die zwei unterschiedlichen Erklärungen aufgeführt.

Tabelle 6.5: Überblick über die Ergebnisse der bevorzugten Erklärung in Prozentangabe [Quelle: In Anlehnung an [Bur23]].

Aufgabe	Erklärung 1	Ergebnis	Erklärung 2	Ergebnis
A	Wert BMI	33,8 %	Kategorie BMI	66,2 %
B	Gewicht und Größe	23,1 %	Kategorie BMI	76,9 %
C	Gewicht und Größe	40,8 %	Wert BMI	59,2 %
D	Datum	27,7 %	Jahreszeit	72,3 %
E	Breitengrad	9,2 %	Klimazone	90,8 %

Der Body-Mass-Index wird gegenüber den einfachen Merkmalen Gewicht und Größe bevorzugt, während die Gewichtsklasse gegenüber beiden präferiert wird. Die Jahreszeit wird gegenüber dem genauen Datum bevorzugt und die Klimazone eines Ortes gegenüber dem Breitengrad.

Hypothese 1: Explizite Merkmale sind nachvollziehbarer.

In jeder der fünf Aufgaben wurden die expliziten Merkmale bevorzugt. Die Erklärung, die das explizite Merkmal enthielt, wurde generell häufiger bevorzugt als die Erklärung mit den einfachen Merkmalen. Die Kommentare der Teilnehmer unterstützen die Erkenntnis. 83 % der Teilnehmer nannten als

wichtigen Grund für diese Entscheidung, dass eine Erklärung möglichst einfach zu verstehen sein soll. Weiter stellten die Teilnehmer heraus, dass eine Erklärung kurz sein soll. Ebenso erklärten sie, dass sie eine Erklärung überspringen würden, wenn diese zu lang ist. Zusammenfassend lässt sich daraus ableiten, dass explizite Merkmale, die von den Teilnehmern nachvollzogen werden können, bevorzugt werden, da diese die Erklärung kompakter gestalten.

Hypothese 2: Detaillierte Merkmale werden in kritischen Situationen bevorzugt.

In kritischen Situationen wie bei der medizinischen Aufgabe bevorzugen mehr als 50 % der Teilnehmer die detaillierten Informationen. Die Hypothese wird durch die Kommentare der Versuchsteilnehmer gestützt. So gaben diese an, dass detaillierte Merkmale gewünscht sind, wenn es um die eigene Gesundheit geht. Weiterhin gaben die Teilnehmer an, dass sie noch mehr Informationen erhalten möchten, auch wenn sie diese nicht vollständig verstehen. Im Gegensatz dazu argumentierten die Teilnehmer, dass in der Aufgabe D die Wettervorhersage von derart geringer Bedeutung ist, dass eine Erklärung wenig bis gar nicht detailliert sein muss bzw. keine benötigt wird.

Während insgesamt die detaillierten Merkmale weniger häufig bevorzugt werden, scheint der Schwierigkeitsgrad der Situation einen Einfluss auf diese Präferenz zu haben.

Hypothese 3: Detaillierte Merkmale werden bevorzugt, wenn die expliziten Merkmale leicht zu erschließen sind.

Wie auch bei der ersten Hypothese festgestellt werden konnte, werden detaillierte Merkmale insgesamt weniger bevorzugt. Dies ist jedoch vom Schwierigkeitsgrad abhängig. Der Vergleich der beiden meteorologischen Beispiele zeigt: Sofern ein Merkmal leicht herzuleiten ist, werden die expliziten Merkmale weniger bevorzugt. Wird also davon ausgegangen, dass die meisten Teilnehmer die Zuordnung von Breitengraden zu Klimazonen nicht direkt ableiten können, ist die Herleitung in Aufgabe E für einen Teilnehmer ohne diesbezügliches Vorwissen verhältnismäßig schwer. Daher wählten nur 9,2 % die Detailinformation, die meisten entschieden sich für das explizite Merkmal. Im

Gegensatz dazu ist die Zuordnung des Datums zu einer Jahreszeit einfacher. In dieser Situation wählten 27,7 % die Erklärung, die die detaillierten Merkmale enthielt. Die Kommentare unterstrichen die Hypothese, da die Teilnehmer angaben, dass die Herleitung vom Datum zur Jahreszeit einfach ist, sodass sie diese selber durchführen möchten. Im Gegensatz dazu beschrieben weitere Teilnehmer, dass ihnen das Konzept des Body-Mass-Index nicht geläufig ist (wie auch bei den Breitengraden). Ein wichtiger Faktor bei der Wahl einer Erklärung ist, das Vorwissen der jeweiligen Anwendergruppe zu kennen. Dies ist ein starker Indikator dafür, wie und ob Merkmale (einfach oder explizit) verstanden werden. Auch muss die Darstellung einer Ontologie vom Anwender verstanden werden.

Die Teilnehmer hatten die Möglichkeit, weitere Kommentare abzugeben. Als wichtiger Faktor, den es bei der Erstellung von Erklärungen zu berücksichtigen gibt, wurde von den Teilnehmern das individuelle Vorwissen genannt. Die dritte Hypothese zeigte, dass die Möglichkeit, bestimmte Herleitungen durchzuführen, einen großen Einfluss auf die Art und Weise hat, wie Merkmale wahrgenommen werden. Beispielsweise könnte vorab eine Art Quiz mit der Person durchgeführt werden, für die die Erklärung bestimmt ist, woraus sich ermitteln lässt, wie detailliert die Erklärung tatsächlich sein soll. Einige Teilnehmer schlugen einen interaktiven Ansatz vor. Dabei sollte es möglich sein, Fragen zu einer gegebenen Erklärung zu stellen. Beispielsweise könnte eine Erklärung zunächst in einer kurzen Form gegeben werden, wie in der ersten Hypothese empfohlen. Darauf aufbauend könnten nach Bedarf immer mehr Details gegeben werden, sofern der Anwender daran Interesse hat.

6.5 Zusammenfassung

In diesem Kapitel wurden insgesamt drei durchgeführte Benutzerstudien vorgestellt. Die erste Benutzerstudie wurde mit einem maritimen Experten als Versuchsteilnehmner durchgeführt. Die Aufgabe des Experten bestand darin, eine Schätzung über einen Schiffstyp anhand vorliegender Merkmale wie der

Geschwindigkeit oder der Trajektorie eines Schiffes vorherzusagen. Anschließend bekam der Experte eine Unterstützung in Form einer Vorhersage, die basierend auf einer Blackbox erzeugt wurde, sowie unterschiedliche Arten von Erklärungen. Nachdem der Experte die Unterstützung erhalten hatte, hatte dieser die Möglichkeit, seine erste Schätzung des Schiffstyps anzupassen. Die Ergebnisse zeigten, dass der Experte in 50 % der Fälle seine erste Schätzung revidierte und den Empfehlungen der KI-Unterstützung folgte. Besonders die globale Erklärung und die Darstellung der Merkmalsrelevanz, die mehr über den Entscheidungsprozess des gesamten Modells aussagte, wurden vom Experten bevorzugt. Die Nachvollziehbarkeit der Merkmale innerhalb der Erklärungen wurde vom Versuchsteilnehmer bemängelt, da diese für ihn nicht klar genug definiert waren. Die zweite Benutzerstudie folgte dem Aufbau der ersten Benutzerstudie, jedoch wurde hierfür ein anderer Kontext gewählt. Die Aufgabe bestand darin, die Abschlussnote von Schülern basierend auf unterschiedlichen Merkmalen wie z. B. der familiären Situation oder der Fehlstunden vom Unterricht zu schätzen. Auch für diese Untersuchung erhielten die insgesamt 80 Versuchsteilnehmer eine Unterstützung in Form einer Vorhersage, die entweder basierend auf einer Whitebox oder einer Blackbox erzeugt wurde und je nach Versuchsausprägung eine Art der Erklärung enthielt. Entsprechend den Ergebnissen der ersten Benutzerstudie wurden hierbei als Erklärungen nur noch lokale und globale Erklärungen anhand der Merkmalsrelevanz gewählt. Durch die Studie konnte festgestellt werden, dass alle Teilnehmer ihre Vorhersagen verbesserten, nachdem diese eine KI-Unterstützung erhielten. Weiter war zu beobachten, dass die Teilnehmer, die eine lokale Erklärung erhielten, ihre Schätzungen mehr anpassten als die Teilnehmer, die eine globale Erklärung erhielten.

Insgesamt zeigen beide Benutzerstudien, dass – unabhängig von der Anwendergruppe – Erklärungen die Nachvollziehbarkeit von Blackbox-Modellen steigern können. Für Domänenexperten eigneten sich globale Erklärungen, da diese in der Regel den gesamten Entscheidungsprozess und nicht nur Teilaspekte des Entscheidungsprozesses verstehen wollen. Für Endbenutzer ohne speziell gefordertes Fachwissen eigneten sich lokale Erklärungen, da für diese die vorliegende Entscheidung von Interesse war, jedoch nicht der

gesamte Entscheidungsprozess der Blackbox. Die Darstellung der Merkmalsrelevanz wurde von beiden Anwendergruppen gleichermaßen nachvollzogen und präferiert.

Bei beiden Benutzerstudien war die Nachvollziehbarkeit der Merkmale entscheidend. Aus diesem Grund wurde in der dritten Benutzerstudie dieser Sachverhalt mit insgesamt 130 Versuchsteilnehmern genauer untersucht. Hierfür wurden fünf Aufgaben für Teilnehmer, die kein spezielles Fachwissen hatten, formuliert. Die Aufgaben unterschieden sich anhand der verwendeten Merkmale. Zum einen wurden detaillierte Merkmale verwendet, die direkt aus dem Datensatz entnommen wurden, und zum anderen wurden explizite Merkmale verwendet. Die expliziten Merkmale wurden durch die Kombination bestimmter Merkmale erzeugt (z. B. Body-Mass-Index aus Größe und Gewicht). Die Aufgaben unterschieden sich nach dem Schwierigkeitsgrad der Situation und der Komplexität der expliziten Merkmale. Dabei wurden beispielsweise medizinische Situationen kritischer als meteorologische Situationen eingestuft. Die Untersuchungen machten deutlich, dass explizite Merkmale bevorzugt wurden, da diese die Länge einer Erklärung verkürzten. Die detaillierten Merkmale wurden in kritischen Situationen (wie z. B. im medizinischen Bereich) bevorzugt. Außerdem stellte sich heraus, dass die detaillierten Merkmale bevorzugt wurden, wenn die expliziten Merkmale leicht zu erschließen sind.

7 Zusammenfassung und Ausblick

7.1 Zusammenfassung

In der vorliegenden Arbeit wurde ein Vorgehensmodell entworfen, das aufzeigt, welche Arten von Erklärungen im Bereich überwachter maschineller Lernverfahren vorwiegend für den Bereich der Klassifikation extrahiert werden können. Insgesamt werden fünf Arten von Erklärungen unterschieden und definiert: interpretierbare (Whitebox-)Modelle , lokale Surrogate, globale Surrogate, direkte lokale Erklärungen und direkte globale Erklärungen. Diese Formalisierung war bisher in der Literatur so nicht existent. Auf Basis des Vorgehensmodells wurden zwei neuartige Verfahren zur Erzeugung globaler und lokaler Surrogate konzipiert. Beide Verfahren erzeugen die Ergebnisse auf Basis einer Blackbox als Hauptmodell. Das Verfahren zur Erzeugung eines globalen Surrogates erzeugt eine globale Erklärung und versucht somit, die Ergebnisse des gesamten Modells nachvollziehbar zu gestalten. Die Lösung zur Erzeugung eines globalen Surrogates basiert auf der Idee der Regularisierung. Dabei wurde ein neuartiger Regularisierungsbegriff eingeführt, der auf der Idee beruht, häufig vorkommende Muster in Regeln zusammenzufassen. Im Gegensatz zu anderen Lösungen in der Literatur muss das Surrogat-Modell nicht mehrmals nachtrainiert werden, sodass das Verfahren eine stark verkürzte Laufzeit hat. Die Ergebnisse zeigten, dass das Verfahren mit den Referenzverfahren aus der Literatur konkurrieren kann und teilweise genauere Ergebnisse erzeugt. Die Lösung zur Erzeugung eines lokalen Surrogates wird ausgehend von einer kontrafaktischen Instanz gefunden und erzeugt eine Erklärung in der Umgebung einer Entscheidungsgrenze. Dabei wird ein

einzelnes Ergebnis einer Blackbox für den Anwender nachvollziehbar gestaltet. Die Ergebnisse machten deutlich, dass das Verfahren eine höhere Modelltreue verglichen mit dem Referenzverfahren aus der Literatur generiert. Zudem zeigte sich, dass das lokale Surrogat, das als Entscheidungsbaum trainiert wurde, häufiger eine höhere Modelltreue hatte als das Surrogat, das als lineares Modell trainiert wurde. Abschließend wurde ein bislang wenig beachteter Aspekt im Bereich der Erklärbarkeit, der Nachvollziehbarkeit von Erklärungen durch Anwender, im Rahmen von Benutzerstudien untersucht. Insgesamt wurden drei Benutzerstudien durchgeführt, die die Nachvollziehbarkeit von Erklärungen anhand unterschiedlicher Aspekte untersuchten. Die ersten beiden Benutzerstudien gingen der Frage nach, inwieweit lokale und globale Erklärungen beim Verständnis unterstützen können. Die Ergebnisse zeigten, dass globale Erklärungen für Domänenexperten geeignet sind, da diese das gesamte Modell nachvollziehen wollen. Lokale Erklärungen waren dem Domänenexperten nicht ausreichend genug. Im Gegensatz dazu waren für Endanwender ohne Fachwissen über eine Domäne die lokalen Erklärungen ausreichend. Diese bevorzugten nicht, das gesamte Modell zu verstehen, sondern lediglich die Entscheidung, die ihnen vorlag und die sie betraf. Die Darstellung der Merkmalsrelevanz wurde von beiden Anwendergruppen als hilfreich eingeordnet. Ein wichtiger Aspekt bei Erklärungen liegt im Verständnis deren Merkmale. Aus diesem Grund wurde dieser Aspekt einzeln in der dritten Benutzerstudie untersucht.

Die dritte Benutzerstudie untersuchte den Einfluss von expliziten Merkmalen, die bspw. basierend auf einer Ontologie konzipiert werden können. Die expliziten Merkmale wurden aus den ursprünglichen Merkmalen eines Datensatzes erstellt. So kann z. B. aus den ursprünglichen Merkmalen *Größe* und *Gewicht* das explizite Merkmal *BMI* erzeugt werden. Die Untersuchungen ergaben, dass explizite Merkmale insgesamt nachvollziehbarer für die Anwender waren. Jedoch wurden die Ursprungsmerkmale in Situationen wie in medizinischen Belangen bevorzugt, da dort die detaillierten Informationen als wichtig erachtet werden. Außerdem zeigte sich, dass die Ursprungsmerkmale bevorzugt wurden, wenn die expliziten Merkmale vom Anwender selbst leicht zu erschließen waren. Insgesamt machen die Untersuchungen deutlich, dass

unabhängig von der Anwendergruppe Erklärungen die Nachvollziehbarkeit von Vorhersagen eines Blackbox-Modells steigern.

7.2 Ausblick

Das Vorgehensmodell wurde für den Teilbereich des überwachten maschinellen Lernens entworfen. Zukünftige Arbeiten können die Übertragbarkeit auf weitere Lernparadigmen des maschinellen Lernens untersuchen wie z. B. im Bereich des bestärkenden oder des unüberwachten Lernens. Dabei kann analysiert werden, inwieweit die Teilkonzepte daraus übertragen werden können. Somit könnte das Vorgehensmodell schrittweise alle weiteren Bereiche des maschinellen Lernens abdecken, sodass für jedes Paradigma bestimmte Erklärarten definiert werden. Die Regularisierung von neuronalen Netzen zeigt einen vielversprechenden Ansatz. Das Konzept kann in zukünftigen Arbeiten auf weitere Typen von neuronalen Netzen (z. B. Convolutional Neuronal Networks (CNNs) oder Recurrent Neural Networks (RNNs)) übertragen und untersucht werden. In diesem Rahmen kann verglichen werden, wie die Regularisierungen auf unterschiedlichen Arten von Netzen funktioniert. Das Vorgehen des lokalen Surrogat-Modells nimmt an, dass die Entscheidungsgrenze innerhalb einer Hypersphäre vorliegt. Weiter könnte hier überprüft werden, inwieweit die Annahme der Hypersphäre durch eine Hyperebene optimierte Ergebnisse erzeugt. Der Themenbereich der Nutzerstudien bietet ebenso viel Raum für Forschungsarbeiten. Zum einen ist es für die Gestaltung und die Durchführung einheitlicher Benutzerstudien notwendig, ein Werkzeug für die Evaluierung von Erklärungen zu entwerfen, das ein standardisiertes Vorgehen bietet. Neben der Einbindung unterschiedlicher Erklärarten für bestimmte Aufgaben sollen automatisiert Metriken abgeleitet werden. Der Entwurf von Metriken, die Aussagen über eine bestimmte Erklärart geben können, ist essenziell, um Erklärungen zügig und vor allem einheitlich vergleichen zu können. Überdies kann die Nachvollziehbarkeit von Erklärungen anhand anderer Darstellungen wie z. B. durch Heatmaps in Benutzerstudien beleuchtet werden. Hier könnte auch ein interaktiver Ansatz verfolgt werden, sodass Anwendern nach Bedarf mehrere Arten von Erklärungen im Rahmen

eines Dashboards für die Erklärbarkeit dargestellt werden. Im Bereich der Ontologien könnten der Verbund von diesen und Erklärungen näher untersucht werden. Erklärungen können bspw. anhand von Ontologien aufgebaut werden, sodass diese einem bestimmten Aufbau folgen und Datenwerte verständlicher darstellen können.

Literatur

[Ada18] Adadi, A. and Berrada, M.: „Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)“. In: *IEEE access* 6 (2018), S. 52138–52160.

[Adl18] Adler, P.; Falk, C.; Friedler, S. A.; Nix, T.; Rybeck, G.; Scheidegger, C.; Smith, B. and Venkatasubramanian, S.: „Auditing black-box models for indirect influence“. In: *Knowledge and Information Systems* 54.1 (2018), S. 95–122.

[Agr94] Agrawal, R.; Srikant, R. et al.: „Fast algorithms for mining association rules“. In: *Proc. 20th int. conf. very large data bases, VLDB*. Bd. 1215. 1994, S. 487–499.

[Alp19] Alpaydin, E.: Maschinelles Lernen. De Gruyter Oldenbourg, 2019.

[Ama17] Amatriain, X.: More Data or Better Models? Nov. 2017. url: http://technocalifornia.blogspot.de/2012/07/more-data-or-better-models.html.

[And13] Andrzejak, A.; Langner, F. and Zabala, S.: „Interpretable models from distributed data via merging of decision trees“. In: *Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on*. IEEE. 2013, S. 1–9.

[Ann20] Anneken, M.; Strenger, M.; Robert, S. and Beyerer, J.: „Classification of Maritime Vessels using Convolutional Neural Networks“. In: *Artificial Intelligence: Research Impact on Key Industries; the Upper-Rhine Artificial Intelligence Symposium (UR-AI 2020)*. Hrsg. von Christ, A. and Quint, F. 2020, S. 103–114. arXiv: 2010.16241 [cs.AI].

[Ask13] Askham, N.; Cook, D.; Doyle, M.; Fereday, H.; Gibson, M.; Landbeck, U.; Lee, R.; Maynard, C.; Palmerand, G. and Schwarzenbach, J.: „The six primary dimensions for data quality assessment". In: *DAMA UK Working Group* (2013), S. 432–435.

[Bae10] Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K. and M. Zoeller, K.-R.: „How to explain individual classification decisions". In: *Journal of Machine Learning Research* 11.Jun (2010), S. 1803–1831.

[Bal17] Balestriero, R.: „Neural decision trees". In: *arXiv preprint arXiv:1702.07360* (2017).

[Bar09] Barbella, D.; Benzaid, S.; Christensen, J. M.; Jackson, B.; Qin, X. V. and Musicant, D.: „Understanding Support Vector Machine Classifications via a Recommender System-Like Approach." In: *DMIN*. 2009, S. 305–311.

[Bas17] Bastani, O.; Kim, C. and Bastani, H.: „Interpreting blackbox models via model extraction". In: *arXiv preprint arXiv:1705.08504* (2017).

[Ben19] Bengio, Y.: „From system 1 deep learning to system 2 deep learning". In: *Thirty-third Conference on Neural Information Processing Systems*. 2019.

[Ber11] Bertsimas, D.; Chang, A. and Rudin, C.: „Ordered rules for classification: A discrete optimization approach to associative classification". In: *SUBMITTED TO THE ANNALS OF STATISTICS*. Citeseer. 2011.

[Ber53] Berkson, J.: „A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function". In: *Journal of the American Statistical Association* 48.263 (1953), S. 565–599.

[Bia21] Biallass, I.: Der Einsatz von künstlicher Intelligenz in der Justiz. 21. Juli 2021. url: https://www.e-justice-magazin.de/wp-content/uploads/sites/27/2019/07/2_Bialla%C3%9F_e-Justice_01_2019.pdf (besucht am 24. 07. 2019).

[Bib21] Bibal, A.; Lognoul, M.; De Streel, A. and Frénay, B.: „Legal requirements on explainability in machine learning". In: *Artificial Intelligence and Law* 29.2 (2021), S. 149–169.

[Bie11] Bien, J. and Tibshirani, R.: „Prototype selection for interpretable classification". In: *The Annals of Applied Statistics* 5.4 (2011), S. 2403–2424.

[Bir17] Biran, O. and McKeown, K. R.: „Human-Centric Justification of Machine Learning Predictions." In: *IJCAI*. 2017, S. 1461–1467.

[Boh17] Bohanec, M.; Borštnar, M. K. and Robnik-Šikonja, M.: „Explaining machine learning models in sales predictions". In: *Expert Systems with Applications* 71 (2017), S. 416–428.

[Bre01a] Breiman, L.: „Random forests". In: *Machine learning* 45.1 (2001), S. 5–32.

[Bre01b] Breiman, L. et al.: „Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: *Statistical science* 16.3 (2001), S. 199–231.

[Bre17] Breiman, L.: Classification and regression trees. Routledge, 2017.

[Bur19] Burkart, N.; Huber, M. and Faller, P.: „Forcing Interpretability for Deep Neural Networks through Rule-Based Regularization". In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE. 2019, S. 700–705.

[Bur20a] Burkart, N.; Faller, P. M.; Peinsipp, E. and Huber, M. F.: „Batch-wise Regularization of Deep Neural Networks for Interpretability". In: *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE. 2020, S. 216–222.

[Bur20b] Burkart, N.; Huber, M. F. and Anneken, M.: „Supported Decision-Making by Explainable Predictions of Ship Trajectories“. In: *International Workshop on Soft Computing Models in Industrial and Environmental Applications*. Springer. 2020, S. 44–54.

[Bur20c] Burkart, N.; Robert, S. and Huber, M. F.: „Are you sure? Prediction revision in automated decision-making“. In: *Expert Systems* 38.1 (2020), S. 12577.

[Bur21a] Burkart, N.; Franz, M. and Huber, M. F.: „Explanation Framework for Intrusion Detection“. In: *Machine Learning for Cyber Physical Systems*. Springer Vieweg, Berlin, Heidelberg, 2021, S. 83–91.

[Bur21b] Burkart, N. and Huber, M. F.: „A Survey on the Explainability of Supervised Machine Learning“. In: *Journal of Artificial Intelligence Research* 70 (2021), S. 245–317.

[Bur23] Burkart, N.; Steinhauser, J. and Huber, M. F.: „On the effect of explicit features“. In Vorbereitung. 2023.

[Car15] Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M. and Elhadad, N.: „Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission“. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, S. 1721–1730.

[Che15] Chen, D.; Fraiberger, S. P.; Moakler, R. and Provost, F.: „Enhancing transparency and control when drawing data-driven inferences about individuals“. In: *Proceedings of 2016 ICML Workshop on Human Interpretability in Machine Learning*. 2015.

[Che18] Chen, J.; Lécué, F.; Pan, J. Z.; Horrocks, I. and Chen, H.: „Knowledge-based transfer learning explanation“. In: *Sixteenth International Conference on Principles of Knowledge Representation and Reasoning*. 2018.

[Cla89] Clark, P. and Niblett, T.: „The CN2 induction algorithm“. In: *Machine learning* 3.4 (1989), S. 261–283.

[Cla91] Clark, P. and Boswell, R.: „Rule induction with CN2: Some recent improvements“. In: *European Working Session on Learning*. Springer. 1991, S. 151–163.

[Coh95] Cohen, W.: „Fast effective rule induction“. In: *Machine Learning Proceedings 1995*. Elsevier, 1995, S. 115–123.

[Col18] Colonna, J. G.; Gama, J. and Nakamura, E. F.: „A comparison of hierarchical multi-output recognition approaches for anuran classification“. In: *Machine Learning* 107.11 (2018), S. 1651–1671.

[Con19] Confalonieri, R.; Prado, F. M. del; Agramunt, S.; Malagarriga, D.; Faggion, D.; Weyde, T. and Besold, T. R.: „An Ontology-based Approach to Explaining Artificial Neural Networks“. In: *arXiv preprint arXiv:1906.08362* (2019).

[Cor08] Cortez, P. and Silva, A. M. G.: „Using data mining to predict secondary school student performance“. In: EUROSIS-ETI, 2008.

[Cor11] Cortez, P. and Embrechts, M. J.: „Opening black box data mining models using sensitivity analysis“. In: *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on*. IEEE. 2011, S. 341–348.

[Cra96] Craven, M. and Shavlik, J. W.: „Extracting tree-structured representations of trained networks“. In: *Advances in neural information processing systems*. Bd. 8. 1996, S. 24–30.

[Dat16] Datta, A.; Sen, S. and Zick, Y.: „Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems“. In: *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE. 2016, S. 598–617.

[Del20] Deloitte: KI-Studie 2020: Wie nutzen Unternehmen Künstliche Intelligenz? Letzter Abruf: 05.05.2021. 2020. url: https://www2.deloitte.com/de/de/pages/technology-media-and-telecommunications/articles/ki-studie-2020.html.

[Dem13] Demšar, J. et al.: „Orange: Data Mining Toolbox in Python“. In: *Journal of Machine Learning Research* 14 (2013), S. 2349–2353.

[Doa04] Doan, A.; Madhavan, J.; Domingos, P. and Halevy, A.: „Ontology Matching: A Machine Learning Approach“. In: *Handbook on Ontologies*. Hrsg. von Staab, S. and Studer, R. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, S. 385–403.

[Dos17] Doshi-Velez, F. and Kim, B.: „Towards a rigorous science of interpretable machine learning“. In: *arXiv preprint arXiv:1702.08608* (2017).

[Du 09] Du Prel, J.-B.; Hommel, G.; Röhrig, B. and Blettner, M.: „Konfidenzintervall oder p-Wert“. In: *Deutsches Aerzteblatt* 106 (2009), S. 335–339.

[Ede12] Eder, J. S.: „Knowledge graph based search system“. US Patent App. 13/404,109. Juni 2012.

[Efr04] Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. et al.: „Least angle regression“. In: *The Annals of statistics* 32.2 (2004), S. 407–499.

[Ehr16] Ehrlinger, L. and Wöß, W.: „Towards a Definition of Knowledge Graphs.“ In: *SEMANTiCS (Posters, Demos, SuCCESS)* 48 (2016), S. 1–4.

[Elb19] Elbekri, N.; Kling, J. and Huber, M. F.: „A study on trust in black box models and post-hoc explanations“. In: *International Workshop on Soft Computing Models in Industrial and Environmental Applications*. Springer. 2019, S. 35–46.

[Fal19] Faller, P. M.: „Rule Regularization for Global Surrogate Model Fitting“. Bachelor’s Thesis. Karlsruhe Institute of Technology, 2019.

[Fic18] Fico: FICO Explainable Machine Learning Challenge. Juni 2018. url: https://community.fico.com/s/explainable-machine-learning-challenge (besucht am 20. 07. 2021).

[Fis36] Fisher, R. A.: „The use of multiple measurements in taxonomic problems“. In: *Annals of eugenics* 7.2 (1936), S. 179–188.

[Fis90] Fischer, G.; Mastaglio, T.; Reeves, B. and Rieman, J.: „Minimalist explanations in knowledge-based systems“. In: *Twenty-Third Annual Hawaii International Conference on System Sciences*. Bd. 3. 1990, 309–317 vol.3.

[Fla12] Flach, P.: Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press, 2012.

[Fre14] Freitas, A.: „Comprehensible classification models: a position paper“. In: *ACM SIGKDD explorations newsletter*. Bd. 15. 1. ACM, 2014, S. 1–10.

[Fri97] Friedman, N.; Geiger, D. and Goldszmidt, M.: „Bayesian network classifiers“. In: *Machine learning* (1997).

[Frö09] Fröhlich, M. and Pieter, A.: „Cohen's Effektstärken als Mass der Bewertung von praktischer Relevanz–Implikationen für die Praxis“. In: *Schweizerische Zeitschrift für Sportmedizin und Sporttraumatologie* 57.4 (2009), S. 139–142.

[Für94] Fürnkranz, J. and Widmer, G.: „Incremental reduced error pruning“. In: *Machine Learning Proceedings 1994*. Elsevier, 1994, S. 70–77.

[Gan10] Ganchev, K.; Graça, J.; Gillenwater, J. and Taskar, B.: „Posterior regularization for structured latent variable models“. In: *The Journal of Machine Learning Research* 11 (2010), S. 2001–2049.

[Gen19] Geng, Y.; Chen, J.; Jimenez-Ruiz, E. and Chen, H.: Human-centric Transfer Learning Explanation via Knowledge Graph [Extended Abstract]. 2019. arXiv: 1901.08547 [cs.LG].

[Gka16] Gkatzia, D.; Lemon, O. and Rieser, V.: „Natural language generation enhances human decision-making with uncertain information“. In: *arXiv preprint arXiv:1606.03254* (2016).

[Gol15] Goldstein, A.; Kapelner, A.; Bleich, J. and Pitkin, E.: „Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation“. In: *Journal of Computational and Graphical Statistics* 24.1 (2015), S. 44–65.

[Goo16] Goodfellow, I.; Bengio, Y. and Courville, A.: Deep learning book. Bd. 521. 7553. 2016, S. 800.

[Goo17] Goodman, B. and Flaxman, S.: „European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"". In: *AI Magazine* 38.3 (2017), S. 50–57.

[Gru93] Gruber, T. R.: „A translation approach to portable ontology specifications". In: *Knowledge acquisition* 5.2 (1993), S. 199–220.

[Gud17] Gudivada, V.; Apon, A. and Ding, J.: „Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations". In: *International Journal on Advances in Software* 10.1 (2017), S. 1–20.

[Gui18a] Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F. and Pedreschi, D.: „A Survey of Methods for Explaining Black Box Models". In: *ACM Comput. Surv.* 51.5 (2018), 93:1–93:42.

[Gui18b] Guidotti, R.; Monreale, A.; Ruggieri, S.; Pedreschi, D.; Turini, F. and Giannotti, F.: „Local rule-based explanations of black box decision systems". In: *arXiv preprint arXiv:1805.10820* (2018).

[Gun17] Gunning, D.: „Explainable artificial intelligence (xai)". In: *Defense Advanced Research Projects Agency (DARPA)* 2.2 (2017).

[Gun19] Gunning, D. and Aha, D. W.: „DARPA's explainable artificial intelligence program". In: *AI Magazine* 40.2 (2019), S. 44–58.

[Hal17a] Hall, P.; Gill, N.; Kurka, M. and Phan, W.: „Machine Learning Interpretability with H2O Driverless AI". In: *H2O.ai* (2017).

[Hal17b] Hall, P.; Phan, W. and Ambati, S.: Ideas on interpreting machine learning. März 2017. url: https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning.

[Har16] Hara, S. and Hayashi, K.: „Making tree ensembles interpretable". In: *arXiv preprint arXiv:1606.05390* (2016).

[Hel16] Helfert, M. and Ge, M.: „Big data quality-towards an explanation model in a smart city context“. In: *Proceedings of 21st International Conference on Information Quality, Ciudad Real, Spain.* 2016.

[Hen14] Henelius, A.; Puolamäki, K.; Boström, H.; Asker, L. and Papapetrou, P.: „A peek into the black box: exploring classifiers by randomization“. In: *Data mining and knowledge discovery* 28.5-6 (2014), S. 1503–1529.

[Hen16] Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B. and Darrell, T.: „Generating visual explanations“. In: *European Conference on Computer Vision*. Springer. 2016, S. 3–19.

[Hen17] Henelius, A.; Puolamäki, K. and Ukkonen, A.: „Interpreting Classifiers through Attribute Interactions in Datasets“. In: *2017 ICML Workshop on Human Interpretability in Machine Learning (WHI)*. 2017.

[HHI21] HHI, F.: Explainable AI Demos. 21. Juli 2021. url: https://lrpserver.hhi.fraunhofer.de/image-classification (besucht am 22. 07. 2021).

[Hin15] Hinton, G.; Vinyals, O. and Dean, J.: „Distilling the knowledge in a neural network“. In: *arXiv preprint arXiv:1503.02531* (2015).

[Hin17] Hinton, G. and Frosst, N.: „Distilling a Neural Network Into a Soft Decision Tree“. In: *Comprehensibility and Explanation in AI and ML (CEX), AI*IA*. 2017.

[Hoe18] Hoeren, T. and Niehoff, M.: „KI und Datenschutz–Begründungserfordernisse automatisierter Entscheidungen“. In: *RW Rechtswissenschaft* 9.1 (2018), S. 47–66.

[Hof18] Hoffman, R. R.; Mueller, S. T.; Klein, G. and Litman, J.: „Metrics for explainable AI: Challenges and prospects“. In: *arXiv preprint arXiv:1812.04608* (2018).

[Hol16] Holzinger, A.: „Interactive machine learning for health informatics: when do we need the human-in-the-loop?“ In: *Brain Informatics* 3.2 (2016), S. 119–131.

[Hol19] Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K. and Müller, H.: „Causability and explainabilty of artificial intelligence in medicine“. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2019), e1312.

[Hol20] Holzinger, A. and Müller, H.: „Verbinden von Natürlicher und Künstlicher Intelligenz: eine experimentelle Testumgebung für Explainable AI (xAI)“. In: Bd. 57. 1. Springer, 2020, S. 33–45.

[Hol93] Holte, R. C.: „Very simple classification rules perform well on most commonly used datasets“. In: *Machine learning* 11.1 (1993), S. 63–90.

[Hoy16] Hoyt, R. E.; Snider, D.; Thompson, C. and Mantravadi, S.: „IBM Watson Analytics: Automating Visualization, Descriptive, and Predictive Statistics“. In: *JMIR Public Health Surveill* 2.2 (Okt. 2016), e157.

[Hu16] Hu, Z.; Ma, X.; Liu, Z.; Hovy, E. and Xing, E.: „Harnessing deep neural networks with logic rules“. In: *arXiv preprint arXiv:1603.06318* (2016).

[Huy11] Huysmans, J.; Dejaeger, K.; Mues, C.; Vanthienen, J. and Baesens, B.: „An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models“. In: *Decision Support Systems* 51.1 (2011), S. 141–154.

[Jia02] Jiang, T. and Owen, A. B.: „Quasi-regression for visualization and interpretation of black box functions“. In: Stanford University, Stanford, 2002. url: https://pdfs.semanticscholar.org/92d0/1110d9d365d16f619fb303932bee3274ba8f.pdf.

[Kag17] Kaggle: The state of data Science and Machine Learning. Okt. 2017. url: https://www.kaggle.com/surveys/2017.

[Kas88] Kass, R. and Finin, T.: „The Need for User Models in Generating Expert System Explanations“. In: *International Journal of Expert Systems* 1.4 (Okt. 1988).

[Kel08] Keller, K.: Netzbasiertes Lehren und Lernen in der betrieblichen Weiterbildung. Springer, 2008.

[Kes20] Kessler, C.: „KI und Legal Tech. Utopie, Dystopie, Realität“. In: *Digitalisierung, Automatisierung, KI und Recht.* Nomos Verlagsgesellschaft mbH & Co. KG. 2020, S. 605–628.

[Kim14] Kim, B.; Rudin, C. and Shah, J. A.: „The bayesian case model: A generative approach for case-based reasoning and prototype classification“. In: *Advances in Neural Information Processing Systems.* 2014, S. 1952–1960.

[Kim16] Kim, B.; Khanna, R. and Koyejo, O. O.: „Examples are not enough, learn to criticize! Criticism for interpretability“. In: *Advances in Neural Information Processing Systems.* 2016, S. 2280–2288.

[Koc13] Koch, K.-R.: Einführung in die Bayes-Statistik. Springer-Verlag, 2013.

[Kod94] Kodratoff, Y.: „The comprehensibility manifesto“. In: *KDD Nugget Newsletter* 94.9 (1994).

[Koh17] Koh, P. W. and Liang, P.: „Understanding black-box predictions via influence functions“. In: *arXiv preprint arXiv:1703.04730* (2017).

[Koh20] Kohler, K. and Scharte, B.: „Der Einsatz von KI im Bevölkerungsschutz“. In: *CSS Analysen zur Sicherheitspolitik* 260 (2020).

[Kom19] Kommission, E.: „Communication: Building Trust in Human Centric Artificial Intelligence)“. In: *COM* 168 (2019), S. 1–11.

[Kos93] Koschnick, W. J.: „Standardwörterbuch für die Sozialwissenschaften/Bd. 2. Deutsch-Englisch Teil 2 M-Z“. In: *Standardwörterbuch für die Sozialwissenschaften= Standard dictionary of the social sciences* (1993).

[Kra16] Krause, J.; Perer, A. and Ng, K.: „Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models“. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM. 2016, S. 5686–5697.

[Kra19] Krause, M. and Natterer, E.: „Maschinelles Lernen“. In: *Wie Maschinen lernen*. Springer, 2019, S. 21–27.

[Kup11] Kuper, H.: „Quantitative Auswertungsverfahren“. In: *Empirische Bildungsforschung*. Springer, 2011, S. 121–129.

[Kur20] Kurt, N.: „Hypothesentests“. In: *Stochastik für Informatiker*. Springer, 2020, S. 145–163.

[Lag19] Lage, I.; Chen, E.; He, J.; Narayanan, M.; Kim, B.; Gershman, S. and Doshi-Velez, F.: „An evaluation of the human-interpretability of explanation“. In: *arXiv preprint arXiv:1902.00006* (2019).

[Lak16] Lakkaraju, H.; Bach, S. H. and Leskovec, J.: „Interpretable decision sets: A joint framework for description and prediction“. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, S. 1675–1684.

[Lak17] Lakkaraju, H.; Kamar, E.; Caruana, R. and Leskovec, J.: „Interpretable & explorable approximations of black box models“. In: *arXiv preprint arXiv:1707.01154* (2017).

[Lak19] Lakkaraju, H.; Kamar, E.; Caruana, R. and Leskovec, J.: „Faithful and Customizable Explanations of Black Box Models“. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019.

[Lau18] Laugel, T.; Renard, X.; Lesot, M.-J.; Marsala, C. and Detyniecki, M.: „Defining locality for surrogates in post-hoc interpretablity“. In: *arXiv preprint arXiv:1806.07498* (2018).

[Lec20] Lecue, F.: „On the role of knowledge graphs in explainable AI“. In: *Semantic Web* 11.1 (2020), S. 41–51.

[Lei18] Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R. J. and Wasserman, L.: „Distribution-Free Predictive Inference for Regression". In: *Journal of the American Statistical Association* 113.523 (2018), S. 1094–1111.

[Let12] Letham, B.; Rudin, C.; McCormick, T. H. and Madigan, D.: „Building interpretable classifiers with rules using Bayesian analysis". In: *Department of Statistics Technical Report tr609, University of Washington* 9.3 (2012), S. 1350–1371.

[Let15] Letham, B.; Rudin, C.; McCormick, T. H.; Madigan, D. et al.: „Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model". In: *The Annals of Applied Statistics* 9.3 (2015), S. 1350–1371.

[Lev18] Levin, S. and Wong, J. C.: „Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian". In: *The Guardian* 19 (2018).

[Lip18] Lipton, Z. C.: „The mythos of model interpretability". In: *Queue* 16.3 (2018), S. 31–57.

[Lor20] Lorenz, U.: „Grundbegriffe des Bestärkenden Lernens". In: *Reinforcement Learning*. Springer, 2020, S. 13–20.

[Lou12] Lou, Y.; Caruana, R. and Gehrke, J.: „Intelligible models for classification and regression". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012, S. 150–158.

[Lou13] Lou, Y.; Caruana, R.; Gehrke, J. and Hooker, G.: „Accurate intelligible models with pairwise interactions". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2013, S. 623–631.

[Lun17] Lundberg, S. M. and Lee, S.: „A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Hrsg. von Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S. and Garnett, R. Curran Associates, Inc., 2017, S. 4765–4774.

[Mal17] Malioutov, D. M.; Varshney, K. R.; Emad, A. and Dash, S.: „Learning interpretable classification rules with boolean compressed sensing“. In: *Transparent Data Mining for Big and Small Data*. Springer, 2017, S. 95–121.

[Mar07] Martens, D.; Backer, M. D.; Haesen, R.; Vanthienen, J.; Snoeck, M. and Baesens, B.: „Classification with ant colony optimization“. In: *IEEE Transactions on Evolutionary Computation* 11.5 (2007), S. 651–665.

[Mar08] Martens, D.; Huysmans, J.; Setiono, R.; Vanthienen, J. and Baesens, B.: „Rule extraction from support vector machines: an overview of issues and application in credit scoring“. In: *Rule extraction from support vector machines* (2008), S. 33–63.

[Mar09] Martens, D.; Baesens, B. and Gestel, T. V.: „Decompositional rule extraction from support vector machines by active learning“. In: *IEEE Transactions on Knowledge and Data Engineering* 21.2 (2009), S. 178–191.

[Mar11] Martens, D.; Vanthienen, J.; Verbeke, W. and Baesens, B.: „Performance of classification models from a user perspective“. In: *Decision Support Systems* (2011).

[McG07] McGuinness, D. L.; Ding, L.; Silva, P. P. da and Chang, C.: „PML 2: A Modular Explanation Interlingua“. In: *ExaCt*. 2007.

[Mei07] Meier, M.: Intrusion Detection effektiv!: Modellierung und Analyse von Angriffsmustern. Springer-Verlag, 2007.

[Mel18] Melis, D. A. and Jaakkola, T.: „Towards robust interpretability with self-explaining neural networks“. In: *Advances in Neural Information Processing Systems*. 2018.

[Mil19] Miller, T.: „Explanation in artificial intelligence: Insights from the social sciences“. In: *Artificial Intelligence* 267 (2019), S. 1–38.

[Moh12] Mohammed, O.; Benlamri, R. and Fong, S.: „Building a Diseases Symptoms Ontology for Medical Diagnosis: An Integrative Approach“. In: Dez. 2012. doi: 10.1109/FGCT.2012.6476567.

[Mol20] Molnar, C.; Casalicchio, G. and Bischl, B.: „Interpretable Machine Learning–A Brief History, State-of-the-Art and Challenges“. In: *arXiv preprint arXiv:2010.09337* (2020).

[Mon17] Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W. and Müller, K. R.: „Explaining nonlinear classification decisions with deep taylor decomposition“. In: *Pattern Recognition* (2017).

[Mur19] Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R. and Yu, B.: „Interpretable machine learning: definitions, methods, and applications“. In: *arXiv preprint arXiv:1901.04592* (2019).

[Nar18] Narkhede, S.: „Understanding auc-roc curve“. In: *Towards Data Science* 26 (2018), S. 220–227.

[Nel65] Nelder, J. A. and Mead, R.: „A simplex method for function minimization“. In: *The computer journal* 7.4 (1965), S. 308–313.

[Ote16] Otero, F. E. B. and Freitas, A.: „Improving the Interpretability of Classification Rules Discovered by an Ant Colony Algorithm: Extended Results“. In: *Evolutionary Computation* 24.3 (2016), S. 385–409.

[Pan20] Panigutti, C.; Perotti, A. and Pedreschi, D.: „Doctor XAI: an ontology-based approach to black-box sequential data classification explanations“. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, S. 629–639.

[Phi18] Phillips, R.; Chang, K. H. and Friedler, S. A.: „Interpretable active learning“. In: *Conference on fairness, accountability and transparency*. PMLR. 2018, S. 49–61.

[Plu18] Plumb, G.; Molitor, D. and Talwalkar, A. S.: „Model agnostic supervised local explanations“. In: *Advances in Neural Information Processing Systems* 31 (2018), S. 2515–2524.

[Pol18] Poland, K.; McKay, M. P.; Bruce, D. and Becic, E.: „Fatal crash between a car operating with automated control systems and a tractor-semitrailer truck“. In: *Traffic injury prevention* 19.sup2 (2018), S153–S156.

[Pou21] Pourssabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Wortman Vaughan, J. W. and Wallach, H.: „Manipulating and measuring model interpretability“. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, S. 1–52.

[Pub18] Publio, G. C.; Esteves, D.; Ławrynowicz, A.; Panov, P.; Soldatova, L.; Soru, T.; Vanschoren, J. and Zafar, H.: „ML-Schema: Exposing the Semantics of Machine Learning with Schemas and Ontologies“. In: *arXiv preprint arXiv:1807.05351* (2018).

[Qui14] Quinlan, J. R.: C4.5: programs for machine learning. Elsevier, 2014.

[Qui86] Quinlan, J. R.: „Induction of decision trees“. In: *Machine learning* 1.1 (1986), S. 81–106.

[Qui90] Quinlan, J. R.: „Learning logical definitions from relations“. In: *Machine learning* 5.3 (1990), S. 239–266.

[Ras20] Rasley, J.; Rajbhandari, S.; Ruwase, O. and He, Y.: „Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters“. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, S. 3505–3506.

[Rea93] Read, S. J. and Marcus-Newhall, A.: „Explanatory coherence in social explanations: A parallel distributed processing account.“ In: *Journal of Personality and Social Psychology* 65.3 (1993), S. 429.

[Rib16a] Ribeiro, M. T.; Singh, S. and Guestrin, C.: „Model-agnostic interpretability of machine learning“. In: *arXiv preprint arXiv:1606.05386* (2016).

[Rib16b] Ribeiro, M. T.; Singh, S. and Guestrin, C.: „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*. ACM. 2016, S. 1135–1144.

[Rib18] Ribeiro, M. T.; Singh, S. and Guestrin, C.: „Anchors: High-Precision Model-Agnostic Explanations“. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*. 2018.

[Rob08] Robnik-Šikonja, M. and Kononenko, I.: „Explaining classifications for individual instances“. In: *IEEE Transactions on Knowledge and Data Engineering* 20.5 (2008), S. 589–600.

[Ros17] Ross, A. S.; Hughes, M. C. and Doshi-Velez, F.: „Right for the right reasons: Training differentiable models by constraining their explanations“. In: *arXiv preprint arXiv:1703.03717* (2017).

[Ros18] Ross, C. and Swetlitz, I.: „IBM's Watson supercomputer recommended 'unsafe and incorrect'cancer treatments, internal documents show“. In: *Stat* 25 (2018).

[Ros20] Roscher, R.; Bohn, B.; Duarte, M. F. and Garcke, J.: „Explainable machine learning for scientific insights and discoveries“. In: *IEEE Access* 8 (2020), S. 42200–42216.

[Rud18] Rudin, C.: „Please Stop Explaining Black Box Models for High Stakes Decisions“. In: *arXiv preprint arXiv:1811.10154* (2018).

[Rüp05] Rüping, S.: „Learning with local models“. In: *Local Pattern Detection*. Springer, 2005, S. 153–170.

[Rüp06] Rüping, S.: „Learning interpretable models“. In: *Doctoral Dissertation*. University of Dortmund, 2006.

[Rus02] Russell, S. and Norvig, P.: „Artificial intelligence: a modern approach“. In: 2002.

[Rus03] Rustenbach, S. J.: „Metaanalyse“. In: *Eine anwendungsorientierte Einführung* 1 (2003).

[Sam17] Samek, W.; Wiegand, T. and Müller, K. R.: „Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models“. In: *arXiv preprint arXiv:1708.08296* (2017).

[Sch07] Schetinin, V.; Fieldsend, J. E.; Partridge, D.; Coats, T. J.; Krzanowski, W. J.; Everson, R. M.; Bailey, T. C. and Hernandez, A.: „Confident interpretation of Bayesian decision tree ensembles for clinical applications“. In: *IEEE Transactions on Information Technology in Biomedicine*. Bd. 11. 3. IEEE, 2007, S. 312–319.

[Sch19a] Schaaf, N.; Huber, M. and Maucher, J.: „Enhancing decision tree based interpretation of deep neural networks through l1-orthogonal regularization“. In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE. 2019, S. 42–49.

[Sch19b] Schmidt, P. and Biessmann, F.: „Quantifying Interpretability and Trust in Machine Learning Systems“. In: *arXiv preprint arXiv:1901.08558* (2019).

[Set08] Setiono, R.; Baesens, B. and Mues, C.: „Recursive neural network rule extraction for data with mixed attributes“. In: *IEEE Transactions on Neural Networks* 19.2 (2008), S. 299–307.

[Sha18] Sharafaldin, I.; Lashkari, A. H. and Ghorbani, A. A.: „Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization.“ In: *ICISSP*. 2018, S. 108–116.

[Shr16] Shrikumar, A.; Greenside, P.; Shcherbina, A. and Kundaje, A.: „Not Just a Black Box: Learning Important Features Through Propagating Activation Differences“. In: *33rd International Conference on Machine Learning*. 2016.

[Si13] Si, Z. and Zhu, S. C.: „Learning and-or templates for object recognition and detection“. In: *IEEE transactions on pattern analysis and machine intelligence*. Bd. 35. 9. IEEE, 2013, S. 2189–2205.

[Sil17] Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A. et al.: „Mastering the game of go without human knowledge“. In: *Nature* 550.7676 (2017), S. 354.

[Smi17] Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F. and Wattenberg, M.: „Smoothgrad: removing noise by adding noise“. In: *arXiv preprint arXiv:1706.03825* (2017).

[Smi20] Smith, G.; Mansilla, R. and Goulding, J.: „Model Class Reliance for Random Forests“. In: *Advances in Neural Information Processing Systems* 33 (2020), S. 1–11.

[Štr10] Štrumbelj, E.; Bosnić, Z.; Kononenko, I.; Zakotnik, B. and Kuhar, C.: „Explanation and reliability of prediction models: the case of breast cancer recurrence“. In: *Knowledge and information systems* 24.2 (2010), S. 305–324.

[Su15] Su, G.; Wei, D.; Varshney, K. R. and Malioutov, D. M.: „Interpretable two-level Boolean rule learning for classification“. In: *arXiv preprint arXiv:1511.07361* (2015).

[Su16] Su, G.; Wei, D.; Varshney, K. R. and Malioutov, D. M.: „Learning sparse two-level boolean rules“. In: *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2016, S. 1–6.

[Sub07] Subianto, M. and Siebes, A.: „Understanding discrete classifiers with a case study in gene prediction“. In: *Seventh IEEE International Conference on Data Mining 2007*. IEEE. 2007, S. 661–666.

[Sun16] Sundararajan, M.; Taly, A. and Yan, Q.: „Gradients of counterfactuals“. In: *arXiv preprint arXiv:1611.02639* (2016), S. 1–19.

[Sun20] Sunyaev, A.: „Vertrauenswürdige Systeme mit künstlicher Intelligenz“. In: *Wirtschaftsinformatik & Management* (2020), S. 1–3.

[Swa91] Swartout, W.; Paris, C. and Moore, J.: „Explanations in knowledge systems: design for explainable expert systems“. In: *IEEE Expert* 6.3 (1991), S. 58–64.

[Tav09] Tavallaee, M.; Bagheri, E.; Lu, W. and Ghorbani, A. A.: „A detailed analysis of the KDD CUP 99 data set“. In: *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*. IEEE. 2009, S. 1–6.

[Tet05] Tetreault, B. J.: „Use of the Automatic Identification System (AIS) for maritime domain awareness (MDA)“. In: *Proceedings of OCEANS 2005 MTS/IEEE*. IEEE. 2005, S. 1590–1594.

[Tha89] Thagard, P.: „Explanatory coherence“. In: *Behavioral and brain sciences* 12.3 (1989), S. 435–502.

[Tib96] Tibshirani, R.: „Regression shrinkage and selection via the lasso“. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), S. 267–288.

[Tol17] Tolomei, G.; Silvestri, F.; Haines, A. and Lalmas, M.: „Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking“. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, S. 465–474.

[Tur16] Turner, R.: „A model explanation system“. In: *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2016, S. 1–6.

[Ust14] Ustun, B. and Rudin, C.: „Methods and models for interpretable linear classification“. In: *arXiv preprint arXiv:1405.4047* (2014), S. 1–57.

[Ust16] Ustun, B. and Rudin, C.: „Supersparse linear integer models for optimized medical scoring systems“. In: *Machine Learning* 102.3 (2016), S. 349–391.

[Ust17] Ustun, B. and Rudin, C.: „Optimized risk scores“. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, S. 1125–1134.

[Ved17] Vedantam, R.; Bengio, S.; Murphy, K.; Parikh, D. and Chechik, G.: „Context-aware captions from context-agnostic supervision“. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, S. 251–260.

[Wac17] Wachter, S.; Mittelstadt, B. and Russell, C.: „Counterfactual explanations without opening the black box: Automated decisions and the GDPR“. In: *Harv. JL & Tech.* 31 (2017), S. 841.

[Wan15a] Wang, F. and Rudin, C.: „Falling Rule Lists“. In: *18th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2015, S. 1013–1022.

[Wan15b] Wang, J.; Fujimaki, R. and Motohashi, Y.: „Trading interpretability for accuracy: Oblique treed sparse additive models“. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, S. 1245–1254.

[Wan15c] Wang, T.; Rudin, C.; Doshi-Velez, F.; Liu, Y.; Klampfl, E. and MacNeille, P.: „Or's of and's for interpretable classification, with application to context-aware recommender systems“. In: *arXiv preprint arXiv:1504.07614* (2015).

[Wan16] Wang, T.; Rudin, C.; Velez-Doshi, F.; Liu, Y.; Klampfl, E. and MacNeille, P.: „Bayesian rule sets for interpretable classification“. In: *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE. 2016, S. 1269–1274.

[Wan96] Wang, R. Y. and Strong, D. M.: „Beyond accuracy: What data quality means to data consumers“. In: *Journal of management information systems* 12.4 (1996), S. 5–33.

[Wat19] Watson, D. S.; Krutzinna, J.; Bruce, I. N.; Griffiths, C. E.; McInnes, I. B.; Barnes, M. R. and Floridi, L.: „Clinical applications of machine learning algorithms: beyond the black box“. In: *Bmj* 364 (2019).

[Wel17] Weller, A.: „Challenges for Transparency“. In: *arXiv preprint arXiv:1708.01870* (2017).

[Wu18] Wu, M.; Hughes, M. C.; Parbhoo, S.; Zazzi, M.; Roth, V. and Doshi-Velez, F.: „Beyond sparsity: Tree regularization of deep models for interpretability“. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[Wu19] Wu, M.; Parbhoo, S.; Hughes, M.; Kindle, R.; Celi, L.; Zazzi, M.; Roth, V. and Doshi-Velez, F.: „Regional Tree Regularization for Interpretability in Black Box Models“. In: *arXiv preprint arXiv:1908.04494* (2019).

[Yan17] Yang, H.; Rudin, C. and Seltzer, M.: „Scalable Bayesian rule lists“. In: *International Conference on Machine Learning*. PMLR. 2017, S. 3921–3930.

[Yan18a] Yang, C.; Rangarajan, A. and Ranka, S.: „Global Model Interpretation via Recursive Partitioning“. In: *arXiv preprint arXiv:1802.04253* (2018).

[Yan18b] Yang, Y.; Morillo, I. G. and Hospedales, T. M.: „Deep neural decision trees“. In: (2018), S. 1–7.

[Yin03] Yin, X. and Han, J.: „CPAR: Classification based on predictive association rules“. In: *Proceedings of the 2003 SIAM International Conference on Data Mining*. SIAM. 2003, S. 331–335.

[Zak97] Zaki, M. J.; Parthasarathy, S.; Ogihara, M. and Li, W.: „Parallel algorithms for discovery of association rules“. In: *Data mining and knowledge discovery* 1.4 (1997), S. 343–373.

[Zou05] Zou, H. and Hastie, T.: „Regularization and variable selection via the elastic net“. In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), S. 301–320.

Publikationen

[1] El Bekri, N.; Angele, S. and Peinsipp-Byma, E.: „Classification of short-lived objects using an interactive adaptable assistance system“. In: *Next-Generation Analyst III*. Bd. 9499. International Society for Optics und Photonics. 2015, S. 949908.

[2] El Bekri, N.; Angele, S.; Ruckhäberle, M.; Peinsipp-Byma, E. and Haelke, B.: „RecceMan: an interactive recognition assistance for image-based reconnaissance: synergistic effects of human perception and computational methods for object recognition, identification, and infrastructure analysis“. In: *Target and Background Signatures*. Bd. 9653. International Society for Optics und Photonics. 2015, S. 96530.

[3] El Bekri, N. and Peinsipp-Byma, E.: „Interactive Data Quality Assistance-An Approach for Min (d) ing the Quality of Data“. In: *Proceedings of the International Conference on Data Science (ICDATA)*. The Steering Committee of The World Congress in Computer Science. 2015, S. 62.

[4] El Bekri, N.; Fischer, Y. and Marosz, D.: „A knowledge-based approach for task-oriented mission planning“. In: *Next-Generation Analyst IV*. Bd. 9851. International Society for Optics und Photonics. 2016, S. 98510.

[5] El Bekri, N. and Peinsipp-Byma, E.: „Assuring Data Quality by Placing the User in the Loop“. In: *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE. 2016, S. 468–471.

[6] El Bekri, N.; Peinsipp-Byma, E. and Syndikus, A.: „Cluster Rule Based Algorithm for Detecting Incorrect Data Records“. In: *2016 UKSim-AMSS 18th International Conference on Computer Modelling and Simulation (UKSim)*. IEEE. 2016, S. 67–71.

[7] Burkart, N.; Huber, M. and Faller, P.: „Forcing Interpretability for Deep Neural Networks through Rule-Based Regularization“. In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE. 2019, S. 700–705.

[8] Elbekri, N.; Kling, J. and Huber, M. F.: „A study on trust in black box models and post-hoc explanations“. In: *International Workshop on Soft Computing Models in Industrial and Environmental Applications*. Springer. 2019, S. 35–46.

[9] Burkart, N.; Faller, P. M.; Peinsipp, E. and Huber, M. F.: „Batch-wise Regularization of Deep Neural Networks for Interpretability“. In: *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE. 2020, S. 216–222.

[10] Burkart, N.; Huber, M. F. and Anneken, M.: „Supported Decision-Making by Explainable Predictions of Ship Trajectories“. In: *International Workshop on Soft Computing Models in Industrial and Environmental Applications*. Springer. 2020, S. 44–54.

[11] Burkart, N.; Robert, S. and Huber, M. F.: „Are you sure? Prediction revision in automated decision-making“. In: *Expert Systems* 38.1 (2020), S. 12577.

[12] Anneken, M.; Veerappa, M. and Burkart, N.: „Anomaly Detection and XAI Concepts in Swarm Intelligence“. In: NATO Symposium. 2021.

[13] Becker, M.; Burkart, N.; Birnstill, P. and Beyerer, J.: „A Step towards Global Counterfactual Explanations“. In: Advances in Artificial Intelligence and Machine Learning. 2021.

[14] Burkart, N.; Franz, M. and Huber, M. F.: „Explanation Framework for Intrusion Detection“. In: *Machine Learning for Cyber Physical Systems*. Springer Vieweg, Berlin, Heidelberg, 2021, S. 83–91.

[15] Burkart, N. and Huber, M. F.: „A Survey on the Explainability of Supervised Machine Learning“. In: *Journal of Artificial Intelligence Research* 70 (2021), S. 245–317.

[16] Veerappa, M.; Anneken, M. and Burkart, N.: „Evaluation of Interpretable Association Rule Mining Methods on Time-Series in the Maritime Domain“. In: *Pattern Recognition. ICPR International Workshops and Challenges*. Springer International Publishing. 2021, S. 204–218.

[17] Burkart, N.; Steinhauser, J. and Huber, M. F.: „On the effect of explicit features“. In Vorbereitung. 2023.

Betreute studentische Arbeiten

[1] Fuhrmann, T.: „ Kombination von Clustering und Assoziationsanalyse-Verfahren zur generischen Fehleridentifizierung in Datensätzen“. Bachelor’s Thesis. Karlsruhe Institute of Technology, 2016.

[2] Kockert, L.: „Konzept zur Nutzung von 3D-Daten als Referenzmaterial innerhalb einer Erkennungsunterstützung“. Master’s Thesis. Karlsruhe Institute of Technology, 2016.

[3] Syndikus, A.: „ Entwurf eines Qualitätssicherungsverfahrens auf Basis verschiedener Data-Mining Algorithmen“. Master’s Thesis. Karlsruhe Institute of Technology, 2016.

[4] Dimitrov, D.: „ Hypotheses comparison with attributed graph matching by structural indexing for probabilisitic scene models“. Bachelor’s Thesis. Karlsruhe Institute of Technology, 2017.

[5] Hennebold, C.: „Vergleich modell-spezifischer und modell-agnostischer Verfahren zur Erzeugung von Erklärungen“. Master’s Thesis. Karlsruhe Institute of Technology, 2018.

[6] Kling, J.: „Creating Explainability for Black Box Predictions using Cluster-Representatives’ Feature Importance“. Master’s Thesis. Karlsruhe Institute of Technology, 2018.

[7] Faller, P. M.: „Rule Regularization for Global Surrogate Model Fitting“. Bachelor’s Thesis. Karlsruhe Institute of Technology, 2019.

[8] Helmold, T.: „Functionally-grounded Evaluation of Interpretability and Explainability Ap-proaches in Machine Learning“. Bachelor’s Thesis. Karlsruhe Institute of Technology, 2019.

[9] Veerapa, M.: „Explanation Systemof a Deep Learning Model on Ship Vessel Trajectories“. Master’s Thesis. Hochschule Karlsruhe – University of Applied Sciences (HKA), 2019.

[10] Becker, M. L.: „A Step towards Global Counterfactual Explanations: Approximating the Feature Space through Hierarchical Division and Graph Search“. Master’s Thesis. Karlsruhe Institute of Technology, 2020.

[11] Steinhauser, J.: „Understanding the Decision Space of Black Box Decisions via Ontologies“. Bachelor’s Thesis. Karlsruhe Institute of Technology, 2020.

Abbildungsverzeichnis

Tabellenverzeichnis

Listings