Friedrich-Alexander-Universität
Institute of Information Systems |
WiSo

Lassen sich Verzerrungen in einem tabellarischen Datensatz über ein Machine Learning Modell ermitteln, welches mit diesem Datensatz trainiert wurde?

Gruppe 4, Thema 6

Seminararbeit

im Rahmen des Seminars „Forschungsmethodisches Seminar“

an der Friedrich-Alexander-Universität Erlangen-Nürnberg
am School of Business, Economics and Society
Chair of Digital Industrial Information Systems

Themenersteller:	Prof. Dr. Martin Matzner
Betreuer:	Sven Weinzierl
vorgelegt von:	Rene Jokiel, Philippe Huber 90403 Nürnberg rene.jokiel@fau.de, philippe.huber@fau.de 22811241, 22658871
Abgabetermin:	10. Februar 2023

Abstract

Im Bereich der Entscheidungsfindung werden künstliche Intelligenzen immer interessanter, auch branchen- und bereichsübergreifend. Die Entscheidungen, die eine künstliche Intelligenz trifft, sind sehr stark von den Datensätzen abhängig, mit denen sie trainiert wurde. Aufgrund dessen ist es besonders wichtig, dass ein Testdatensatz integer, repräsentativ und frei von Diskriminierungen ist (Friedler et al., 2019). Unsere Zielsetzung setzt an diesem Problem an, unser Ziel war es, eine Methode in Form eines Artefaktes zu entwicklen, mit der ein Datensatz auf Diskriminierungen überprüft werden kann, indem das Ergebnise eines Entscheidungsbaumes analysiert wird, der mit dem Datensatz trainiert wurde. Dabei wird vor allem gezeigt, was für Verzerrungen im Entscheidungsbaum vorliegen, wenn er mit diesem Datensatz trainiert wird. Demnach lautet unsere Forschungsfrage: "Lassen sich Verzerrungen in einem tabellarischen Datensatz über ein Machine Learning Modell ermitteln, welches mit diesem Datensatz trainiert wurde?". Für die Beantwortung unserer Forschungsfrage haben wir ein Artefakt entwickelt, das wir mit vier selbsterzeugten Datensätzen und zwei Datensätzen vom Python Interpretable Machine Learning GitHub Repository getestet haben. Unsere Forschungsmethode war hierbei Design Science Research (DSR).

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die Bereiche der Künstlichen Intelligenz und des Machine Learnings sind rasant wachsende Forschungsbereiche, welche Ausprägungen in vielen Bereichen vorzuweisen haben. KI-Applikationen und KI-Unterstützung sind weit verbreitet, man findet sie heutzutage in vielen Bereichen, wie in der Medizin, der Produktion, der Lehre, dem Marketing oder auch im Finanz-Sektor. Die Anwendung und Umsetzung von KI-Applikationen sind sehr divers, man verwendet KI-Applikationen beispielsweise für Betrugserkennung (Brotcke, 2022) oder auch in verschiedenen Bereichen der Entscheidungsfindung (Friedler et al., 2019).

Künstliche Intelligenzen (KI) und Maschinelles Lernen (ML) sind und werden immer mehr ein wichtiger Teil in der Entscheidungsfindung, jedoch sind KI, ML und die aus ihnen folgenden Entscheidungen und Entscheidungsunterstützungen stark abhängig von den Trainingsdatensätzen, mit denen sie trainiert werden. Demnach ist es von großer Relevanz für alle Beteiligten, dass innerhalb von Testdatensätzen keinerlei Verzerrungen (Bias) oder Diskriminierungen enthalten sind, da diese vom Algorithmus mitgelernt werden und sich in dem Modell und den Ausgaben der künstlichen Intelligenzen widerspiegeln können. Solche Verzerrungen können zu unfairer Behandlung und Diskriminierung von Gruppen in unserer Gesellschaft führen. So kann es passieren, dass Menschen aufgrund ihrer Ethnie, ihres Geschlechts oder ähnlichen Attributen durch solche Verzerrungen negativer durch die künstliche Intelligenz bewertet werden, was zu direkter Diskriminierung dieser Personengruppen führen kann. Derartige Verzerrungen zu reduzieren sowie Fairness zu schaffen ist dementsprechend ein großes Forschungsfeld geworden (Friedler et al., 2019).

Das Thema dieser Arbeit befindet sich genau im Forschungsfeld der Fair AI. Ziel und Thema der Arbeit ist es, eine Methode in Form eines Artefaktes zu entwerfen und zu entwickeln, welches sich mit solchen Verzerrungen befasst. Das Artefakt soll einen Datensatz auf seine Verzerrungen überprüfen und Ungleichheiten zwischen verschiedenen Personengruppen aufzeigen. Dazu soll ein Entscheidungsbaum verwendet werden, der durch maschinelles Lernen erzeugt wird. Das Artefakt beschränkt sich auf tabellarische Datensätze. Unsere Forschungsfrage lautet hierbei "Lassen sich Verzerrungen in einem tabellarischen Datensatz über ein Machine Learning Modell ermitteln, welches mit diesem Datensatz trainiert wurde?". "Ermitteln" umfasst in diesem Kontext nur das Zeigen, ob es Verzerrungen gibt und wie stark diese sind und nicht, wo diese genau sind oder warum sie existieren. Das Artefakt soll am Ende praktisch ein "Verzerrungs-Prüfer" und kein "Verzerrungs-Finder" sein. Ein wichtiger Teil der Forschungsfrage ist hierbei, wie sehr sich die Verzerrungen eines Datensatzes auf ein Modell maschinellen Lernens übertragen lassen.

2 Literaturrecherche

Um den aktuellen Stand der Forschung zum Thema Fairness beim maschinellen Lernen herauszufinden, haben wir eine Literaturrecherche durchgeführt. Dazu haben wir verschiedene Artikel gelesen, um die für uns relevanten Informationen zu identifizieren. Dabei haben wir uns in unserem Suchstring unter anderem auf Begriffe wie "Fair AI", "Reduce Bias" oder "Ensure Fairness" fokussiert und diese in verschiedenen Kombinationen angegeben. Nachdem uns Scopus und IEEE Xplore insgesamt 206 Treffer lieferten, konnten wir diese durch verschiedene Eliminierungen auf 7 reduzieren und anschließend durch Forward und Backward Search auf 10 erhöhen, wie Tabelle 1 zeigt.

	Scopus	IEEE Xplore	Total
Zurückgelieferte Ergebnisse	131	75	206
Duplikats-Eliminierung	-0	-2	204
Titel-Eliminierung	-115	-66	23
Abstract-Eliminierung	-3	-5	15
Verfügbarkeits-Eliminierung	-4	-0	11
Fulltext-Eliminierung	-3	-1	7
Forward & Backward Search	+1	+2	**10**

Tabelle 1 Reduzierung der Suchergebnisse

Die Auswahl dieser 10 Paper erfolgte nach verschiedenen Kriterien, so haben wir beispielsweise Dokumente ausgeschlossen, die zu sehr auf den mathematischen Aspekt eingehen. Um die gefundenen Paper besser einordnen zu können, haben wir uns Kategorien überlegt und die Artikel in Form einer Konzeptmatrix (Tabelle 2) zusammengetragen. Die gewählten Kriterien umfassen die gewählte Datensatzart, welche Rolle Verzerrungen spielen (geht es eher um die Eliminierung/Minderung, oder darum den Bias zu finden?), und was für Methoden verwendet und angegeben wurden.

	Datensatzart			Verzerrungsumgang (Bias ...)			Methoden		
	Tabellarisch	Nicht Tabellarisch	Nicht Definiert	Eliminierung	Minderung	Bestimmung	Berechnungen	(Pseudo-) Code	Ansätze
Dwork et al. (2012)			X			X	X		
Kamiran and Calders (2012)	X			X		X	X	X	
Roselli et al. (2019)			X		X				
He et al. (2020)	X			X			X		
Mazilu et al. (2020)	X				X	X		X	X
Robert et al. (2020)			X			X			X
Agarwal et al. (2022)		X			X	X		X	X
Beatti et al. (2022)		X			X	X			X
Brotcke (2022)			X			X			
Mehrabi et al. (2022)	X	X			X		X	X	
∑	4	3	4	2	5	7	4	4	4

Tabelle 2 Konzeptmatrix

In Robert et al. (2020) haben wir herausgefunden, dass Fairness in drei Arten unterteilt werden kann: distributiv, prozedural und interaktiv. Distributive Fairness beschreibt, wie fair eine KI bei der Zuteilung von Ressourcen (beispielsweise Gehälter) ist. Prozedurale Fairness bedeutet, dass Prozesse fair und transparent sein sollen. Dies wird dadurch erreicht, dass alle Parteien gehört werden, bevor eine Entscheidung getroffen wird. Für prozedurale Fairness wurden in Robert et al. (2020) sechs Charakteristiken definiert: Konsistenz, unvoreingenommene Verdrängung, Repräsentativität, Korrigierbarkeit, Genauigkeit (Accuracy), sowie ethnische Vertretbarkeit. Bei der interaktiven Fairness geht es darum, wie Mitarbeiter:innen von ihrer Organisation behandelt werden. Die interaktive Fairness kann in zwischenmenschlich und informativ untergliedert werden, wobei ersteres den Respekt und die Würde beschreibt, die Arbeitnehmer:innen entgegengebracht wird, während sich letzteres auf die Transparenz bei Prozessen bezieht.

Auch Roselli et al. (2019) unterteilt Bias in drei Kategorien, jedoch ist hier die Klassifizierung ein wenig anders als bei Robert et al. (2020). Die erste Kategorie umfasst Verzerrungen die entstehen, indem Geschäftsabsichten in die KI-Implementierung einfließen. Die zweite beschreibt solche, die durch die Verteilung der für das Training verwendeten Samples entstehen. In der dritten und letzten Kategorie sind jene Biases enthalten, die in einzelnen Eingabesamples vorhanden sind. Roselli et al. (2019) beschreibt in diesem Artikel eine Reihe von Prozessen, um diese drei Kategorien von Bias zu reduzieren, und geht dabei auf die verschiedenen Probleme ein, die sich in der Verzerrungsreduzierung ergeben.

Eine große Hürde im Bereich der Fair AI ist es, die Accuracy dabei nicht zu stark zu vernachlässigen. Wie schwierig es ist, den idealen Kompromiss zwischen Accuracy und Fairness zu finden, wurde in Kamiran and Calders (2012) erläutert. Dort wird unter anderem eine theoretische Analyse des Trade-Offs durchgeführt. Zudem werden dort auch algorithmische Ansätze präsentiert, wie das Gleichgewicht zwischen Accuracy und Fairness optimiert werden kann.

Mehrabi et al. (2022) befasst sich mit Künstlichen Intelligenzen, die zur Entscheidungsfindung in kritischen Umgebungen wie beispielsweise Krankenhäusern eingesetzt werden. Dazu wurde ein aufmerksamkeitsbasiertes Modell entwickelt, das als Framework für die Klassifizierung verwendet werden kann. Dieses Framework wurde dann dafür benutzt, eine Strategie zur Verzerrungsminderung zu entwickeln. Im Anschluss wurden Experimente mit unterschiedlichen Datensätzen durchgeführt, eines mit tabellarischen und eines mit textuellen Daten.

Viele Algorithmen, die in Bezug auf Fairness optimiert wurden, stehen häufig einem der folgenden drei Probleme gegenüber: (1) Ihre Entscheidungen sind nicht oder nur schwer nachzuvollziehen, (2) lässt die Qualität ihrer Vorhersagen mit der Zeit nach, oder (3) sie sind nicht oder nur schwer auf andere Modelle übertragbar (He et al., 2020). Dafür wurde in He et al. (2020) ein geometrischer Ansatz gewählt, um Korrelationen zwischen Daten und einer Anzahl an geschützten Variablen zu verringern. Die daraus resultierenden Features sind interpretierbar und lassen sich auf andere Modelle übertragen.

In Agarwal et al. (2022) wird versucht, Modellverzerrungen zu reduzieren, indem kontextuell faire Daten aufbereitet werden. Dabei wird ein Algorithmus zur Datenvorverarbeitung gezeigt, welcher dazu den Variationskoeffizienten c_v verwendet und dabei hilft, repräsentative Verzerrungen zu verringern und die Rate der positiven Ergebnisse für eine geschützte Gruppe zu erhöhen. Diese Technik wurde in diesem Artikel primär für Modelle der Objekterkennung und Multi-Label-Bildklassifizierung verwendet.

Aber auch im Bereich der Chatbots wird intensiv geforscht, wie man sie immun gegen Vorurteile machen kann, ohne ihre Lernfähigkeit einzuschränken. So ist in Beattie et al. (2022) vom Chatbot *Taybot* von Microsoft die Rede, der innerhalb von 24 Stunden von einem netten und unterstützenden Helfer zu einem Rassisten wurde, der Nazi-Propaganda verbreitet. Um diesen Effekt zu verhindern, wurden in diesem Projekt zwei Chatbots entwickelt, die jeweils unterschiedliche Ansätze verfolgen: einer verwendet Machine Learning (*Chatterbot*), während der andere (*CodeSpeedy*) Deep Learning nutzt. Dabei stellte sich *CodeSpeedy* sowohl als weniger toxisch heraus als auch zuverlässiger darin, korrekt auf die Konversation einzugehen. Nichtsdestotrotz haben

beide Chatbots ebenfalls toxisches Verhalten angelernt, welches sich jedoch durch Hinzufügen "freundlicher Konversationen" zum Trainingsdatensatz reduzieren liess.

Wie aus Mazilu et al. (2020) hervorgeht, können Datensatzverzerrungen unter anderem in zwei Kategorien unterteilt werden, nämlich (1) solche, die aus ungleicher Repräsentation sensitiver Gruppen hervorgehen, und (2) solche, die durch versteckte Vorurteile durch Proxies für sensible Attribute entstehen. Unter Proxy-Attributen sind solche zu verstehen, die implizit eine sensitive Gruppe repräsentieren. So kann beispielsweise *Nachbarschaft* ein Proxy Attribut für *Rasse* sein.

Mithilfe dieser Quellen konnten wir die Spannweite von Bias im Bereich der KI und dessen Minderung/Eliminierung besser einordnen und uns einen guten Überblick verschaffen, wie der aktuelle Stand der Forschung ist und was daraus bisher hervorging. Es hat sich herausgestellt, dass es sehr viele verschiedene Ansätze gibt, dieses Problem anzugehen und dass viele davon auch erfolgreich sind, auch wenn bisher noch keine perfekte Lösung gefunden wurde.

3 Methode

Da wir in dieser Arbeit ein Artefakt entwickelt haben, fällt sie in die Kategorie der Design Science Research Studien (DSR). Solche Studien sind gemäß Gregor and Hevner (2013) in folgende sieben Abschnitte gegliedert: Einleitung, Literaturrecherche, Methode, Artefaktbeschreibung, Evaluation, Diskussion, und Fazit. An diese Ordnung haben wir uns auch möglichst gehalten.

3.1 Framework zum Wissensbeitrag einer DSR

In Gregor and Hevner (2013) wird ein *DSR Knowledge Contribution Framework* verwendet, welches den Reifegrad der Lösung dem Reifegrad der Anwendungsdomäne gegenüberstellt. Dabei werden Routine Design, Exaptation, Invention und Improvement unterschieden. Routine Design beschreibt die Anwendung bekannter Lösungen auf bekannte Probleme. Dabei findet kein größerer Wissensbeitrag statt. Bei der Exaptation geht es darum, bekannte Lösungen auf neue Probleme zu erweitern. Hier bietet sich sowohl eine Forschungsmöglichkeit als auch ein Wissensbeitrag an. Diese beiden Beiträge sind auch bei der Invention zu finden. Bei dieser ist es das Ziel, neue Lösungen für neue Probleme zu entwickeln. Dabei handelt es sich oft um einen Durchbruch. Unser Artefakt ist der letzten der vier Kategorien zuzuordnen: dem Improvement. Hier wird eine neue Lösung auf ein bekanntes Problem angewandt. Das Ziel davon ist es, existierende Ansätze/Lösungen zu verbessern, indem neue und effizientere Artefakte entwickelt werden. Auch hier finden sich sowohl eine Forschungsmöglichkeit als auch ein Wissensbeitrag.

3.2 Drei Stufen der Beitragsart

Gregor and Hevner (2013) ordnet Artefakte auch in drei verschiedene Stufen der Beitragsart ein. Stufe 1 ist die gegebene Implementierung eines Artefakts, also beispielsweise eine Instanziierung. Da unser Artefakt am ehesten als Methode eingeordnet werden kann, ist es der zweiten Stufe zuzuordnen, nämlich der schöpferischen Designtheorie. Wissen wird dabei als operationales Prinzip oder als Architektur verwendet. Beispiele dafür sind unter anderem Konstrukte, Methoden oder Modelle. Die dritte Stufe umfasst "voll entwickelte Designtheorien über eingebettete Phänomene".

4 Artefaktbeschreibung

Unsere Forschungsfrage "Lassen sich Verzerrungen in einem tabellarischen Datensatz über ein Machine Learning Modell, das mit diesem Datensatz trainiert wurde, ermitteln?", ist, da sie aus dem Forschungsbereich der Fair AI und somit aus dem übergeordneten Forschungsbereich der Artificial Intelligence kommt, technischer Natur, weswegen wir uns einen technischen Lösungsansatz für ihre Bearbeitung und Beantwortung überlegt haben.

4.1 Kerngedanke

Für die Beantwortung der Forschungsfrage müssen wir deskriptive Aussagen über den zu bewertenden Datensatz treffen. Dafür gibt es verschiedene statistische Methoden, wie die Betrachtung von Verteilungen und deren Abständen zueinander innerhalb des Datensatzes und der Berechnung weiterer Kennzahlen basierend auf eben jenen Verteilungen und Abständen. Wir haben uns allerdings für einen anderen Ansatz entschieden und diesen daraufhin auch entwickelt (Dwork et al., 2012). Der Kerngedanke unseres Lösungsansatzes, den wir mittels eines Artefaktes implementiert haben, ist es, sämtliche Verzerrungen, Unfairness und Diskriminierungen, die in einem Datensatz enthalten sind, auf das Modell einer künstlichen Intelligenz zu übertragen und darauf folgend die Ergebnisse der künstlichen Intelligenz zu analysieren, um so Rückschlüsse auf den Ursprungsdatensatz ziehen zu können. In diesem Ansatz wird der Datensatz, der auf Verzerrungen und Fairness geprüft werden soll, selbst gar nicht analysiert, auch nicht die Gewichte innerhalb der künstlichen Intelligenz sondern nur die produzierten Ergebnisse der künstlichen Intelligenz.

4.2 Einschränkungen

Das Artefakt beschränkt sich exklusiv auf tabellarische Datensätze als zu prüfende Datensätze, andere Datensatzarten wie beispielsweise Prozess- oder Bilddaten sind kein valider Input für das Artefakt. Eine weitere Beschränkung bezüglich des Datensatzes ist der Kontext. Es muss in dem Datensatz ein kategorisches Attribut mit genau zwei Ausprägungen enthalten sein, welches den Kontext einer Entscheidung hat. Ein Beispiel für so ein Attribut wäre "WurdeBefördert" mit den Ausprägungen $A_1 = \{Ja, Nein\}$. Des Weiteren müssen die Werte aller Attribute beziehungsweise Spalten auf numerische Werte abgebildet werden können. Die letzte Einschränkung ist, dass es nur zwei verschiedene Arten von Datenarten geben darf, numerische und kategorische. Das bedeutet, dass beispielsweise ein Datensatz, in dem es das Attribut "Bewertung" gibt, in welchem ein ausführlicher Fließtext über die Arbeitsleistung und

das Verhalten des Angestellten am Arbeitsplatz gespeichert wird, nicht geeignet für das Artefakt zum Überprüfen auf Verzerrungen und Diskriminierungen ist. Ganz klassische Attribute, die auch nicht im Datensatz enthalten sein dürfen, sind "Vorname" und "Nachname". Ein Beispiel für ein numerisches Attribut wäre "Gehalt" und ein Beispiel für ein kategorisches Attribut wäre "Geschlecht". Weitere Einschränkungen, auch im Bezug auf den Kontext, gibt es nicht, ein Datensatz kann aus jedem Bereich kommen, solange er diese drei Bedingungen erfüllt.

4.3 Aufbau und Funktionsweise des Artefaktes

Abstrahiert ist das Artefakt aufgebaut wie in Abbildung 1. Es lässt sich hierbei vere-

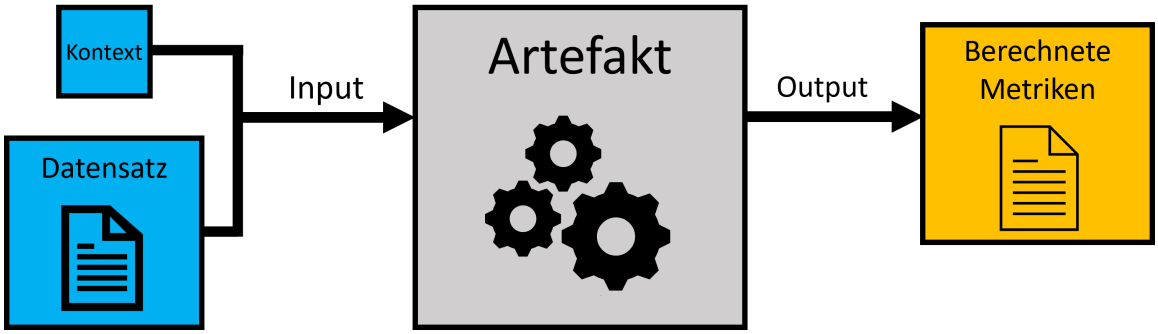

Abbildung 1 Abstrakte Darstellung des Artefaktes

infacht gesagt in drei Seiten unterteilen: die Inputseite, welche von den beschriebenen Einschränkungen betroffen ist, die Prozedurseite beziehungsweise das Artefakt selbst, welche die Prozesse und Funktionen des Artefaktes enthält und die Outputseite, welche das Resultat und die ensprechenden Metriken enthält.

4.3.1 Input

Das Artefakt bekommt zu Beginn zwei Objekte als Input: einen den Einschränkungen entsprechenden Datensatz und ein Kontextobjekt zu dem Datensatz. Das Kontextobjekt enthält wichtige Metainformationen, die benötigt werden, um den Datensatz weiter zu verarbeiten. Im Kontextobjekt sind vier Attribute und eine Methode enthalten, dies kann auch variieren, je nachdem, von welcher Beschaffenheit der Datensatz ist. Die Attribute heißen SpaltenTypZuordnung, WertAbbildungen, RelevanteAttribute und VorherzusagendesAttribut, die Methode heißt BildeWerteAb(). Das Attribut SpaltenTypZuordnung beschreibt, welche Datenarten in welchen Spalten gespeichert werden. Daraus lässt sich also ablesen, welche Spalten numerische Werte und welche Spalten kategorische Werte beinhalten. WertAbbildungen beschreibt, wie kate-

gorische Werte auf numerische Werte abzubilden sind, also welche kategorische Ausprägung welchem numerischen Wert entspricht. Auf dieses Attribut muss die BildeWerteAb() Methode zurückgreifen, wenn sie die kategorischen Werte auf numerische Werte abbildet. Sowohl das Attribut als auch die Methode können weggelassen werden, wenn der Datensatz bereits so transformiert wurde, dass in jeder Zelle nur Zahlen stehen. RelevanteAttribute beschreibt, welche Attribute des Datensatzes für die Prozedur relevant sind, im Normalfall fallen dadurch Attribute wie "Name" oder "ID" weg. Dieses Attribut kann auch weggelassen werden, wenn der Datensatz bereits nur auf relevante Attribute begrenzt wurde. VorherzusagendesAttribut beschreibt, welches Attribut aus dem Datensatz vorhergesagt werden soll, also, welches Attribut das für die Analyse interessante Entscheidungsattribut ist. Das Attribut VorherzusagendesAttribut muss immer in irgendeiner Form vorhanden sein. In der folgenden Abbildung ist das Kontextobjekt in Form einer Klassenkarte visuell zusammengefasst.

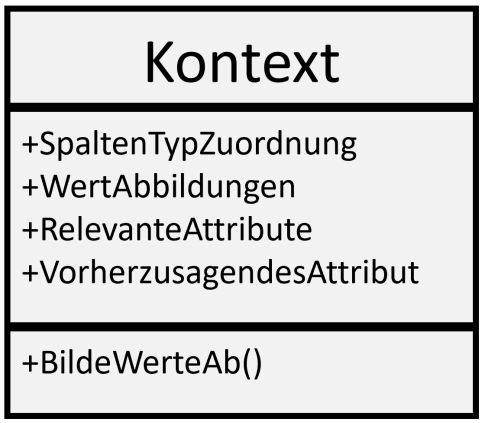

Abbildung 2 Klassenkarte zum Kontextobjekt

Sollte der Datensatz noch nicht passend vorbereitet worden sein, wird das basierend auf den Metainformationen des Kontextobjektes gemacht.

4.3.2 Artefaktprozess

Der Prozess des Artefaktes beginnt damit, dass der Input-Datensatz zufällig in einen Test- und in einen Trainingsdatensatz aufgeteilt wird. Mit dem Trainingsdatensatz wird ein Entscheidungsbaum trainiert, mit dem Testdatensatz wird seine Genauigkeit getestet. Je höher die Genauigkeit des Modells ist, desto aussagekräftiger ist das finale Gesamtergebnis des Artefakts.

Im nächsten Prozessschritt wird ein neuer Datensatz angelegt, welcher auf dem Datensatz basiert und aufbaut, der als Input übergeben wurde. Die Prämisse des neuen Datensatzes ist es, alle Einträge zu enthalten, die es basierend auf den Werten des

Inputdatensatzes geben kann. Zur Erzeugung des Datensatzes wird das Kreuzpodukt aus allen Ausprägungen aller kategorischer Attribute und den Intervallen der numerischen Attribute gebildet und in eine Tabelle eingetragen. Somit erhält man alle möglichen Variationen der Dateneinträge des Input-Datensatzes und noch einige mehr, da bei numerischen Attributen das Intervall zur Berechnung genommen wird anstatt der Ausprägungen. In Abbildung 3 ist die Erzeugung des neuen Datensatzes anhand eines Beispiels visualisiert.

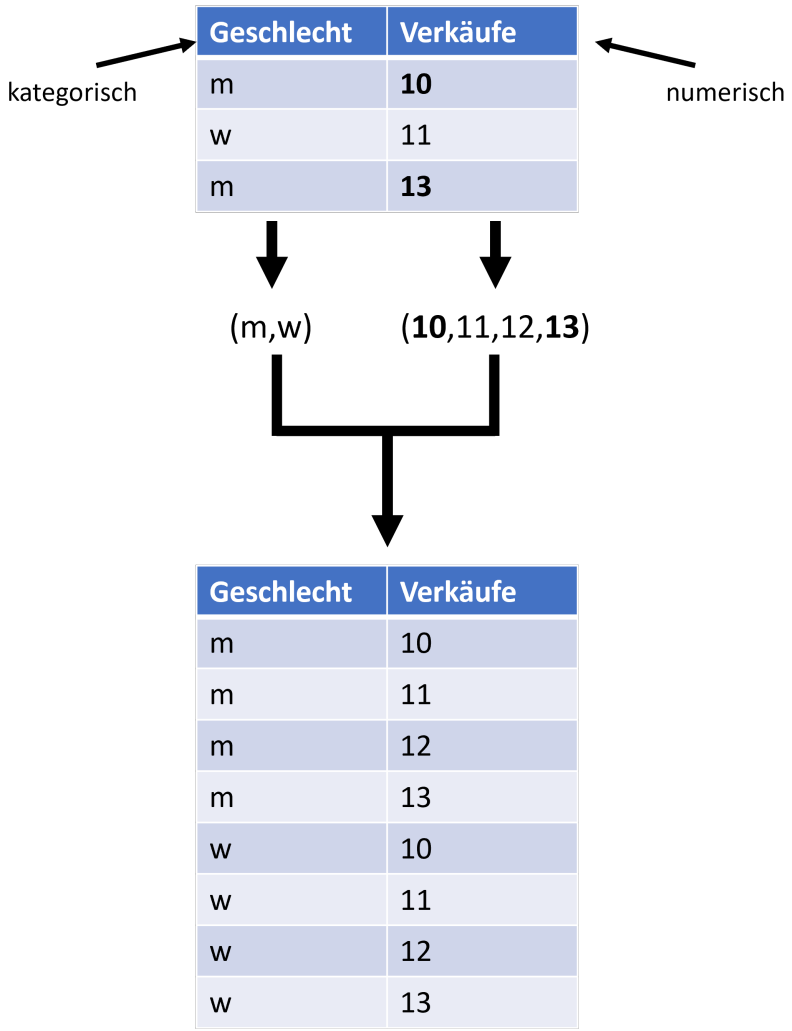

Abbildung 3 Erzeugung neuer Tabelle

In dem Beispiel der Abbildung wurden alle möglichen Personen erzeugt, die es im Rahmen des Ursprungsdatensatzes geben könnte.

Der neu erzeugten Tabelle wird nun eine neue Spalte hinzugefügt, bei welcher es sich um das vorherzusagende Attribut handelt. Die Werte für diese Spalte für die einzelnen Dateneinträge werden von dem mit dem Ursprungsdatensatz trainierten Entscheidungsbaum bestimmt beziehungsweise vorhergesagt.

4.3.3 Output und Outputberechnung

Der letzte Prozessschritt des Artefaktes ist die Erhebung und Berechnung des Outputs. Die neu erzeugte Tabelle wird nun analysiert und auf Verzerrungen und Diskriminierungen untersucht. Für jedes kategorische Attribut wird für jede Ausprägung ermittelt, was die Mindestvorraussetzungen sind, um ein Entscheidungslevel zu erhalten. Da es durch die Einschränkungen nur zwei Ausprägungen, also "0" und "1" beziehungsweise, je nach Kontext, "Nein" und "Ja" gibt, werden die Mindestvorraussetzungen ermittelt, um die Wertunng "1" beziehungsweise "Ja" zu bekommen. Um die Mindestvorraussetzungen zu berechnen wird standardmäßig für jeden Eintrag ein Vergleichswert berechnet, welcher die Summe aller numerischer Attribute ist. Die Ausprägung der kategorischen Attribute eines Dateneintrages gehen dabei nicht mit in die Berechnung ein. Da der Kontext der numerischen Attribute und die Logik hinter ihnen hierfür stark relevant ist, sollte vom Standardverfahren abgewichen werden und eine für den Kontext der Daten passende Formel zur Berechnung der Vergleichswert verwendet werden. Die Mindestvorraussetzung für eine Gruppe ist dann der Dateneintrag der neu generierten Tabelle, der den geringsten Vergleichswert hat und ein "Ja" beziehungsweise eine "1" in der Entscheidungsspalte hat.

Des Weiteren wird für jede Ausprägung der kategorischen Attribute der Anteil an "1"-Entscheidungen berechnet. Angenommen, in dem Beispiel aus der vorherigen Abbildung wäre das Entscheidungsattribut "FürGehaltserhöhungVorschlagen" mit den Ausprägungen "Nein" und "Ja", so würde für das Attribut Geschlecht ermittelt werden, ab wie vielen Verkäufen ein Mann und ab wie vielen Verkäufen eine Frau befördert werden würde. Zudem würde berechnet werden, wie hoch der Anteil an für eine Gehaltserhöhung vorgeschlagenen Männern ist und das Gleiche analog für Frauen. Die Anteile werden in Prozent berechnet und müssen kumuliert 100% ergeben.

Des Weiteren werden Kennzahlen berechnet, die den ganzen Datensatz beschreiben. Diese Kennzahlen wurden extra für den Kontext dieses Projektes erdacht.

Die erste Kennzahl, die berechnet wird, ist der Anforderungsabstand. Berechnet wird der Anforderungsabstand, mit dem Buchstaben B abgekürzt, indem man für jedes kategorische Attribut die Differenz zwischen der Ausprägung mit der höchsten Mindest-

anforderung für die "1"-Entscheidung und der Ausprägung mit der niedrigsten Mindestanforderung berechnet. Diese Differenzen werden aufaddiert und dann durch die Anzahl an kategorischen Attributen geteilt. Die Kennzahl gibt an, wie stark manche Personengruppen im Durchschnitt bevorzugt werden beziehungsweise wie viel mehr diskriminierte und benachteiligte Personengruppen im Durchschnitt leisten oder erbringen müssen, um gleich behandelt zu werden. Als mathematische Formel ausgedrückt sieht der Anforderungsabstand wie folgt aus:

$B = \frac{\sum_{i=0}^{m} max(n_i) - min(n_i)}{m}$

Die Variablen in der Formel sind wie folgt zuzuordnen:

- B: Der Anforderungsabstand
- m: Anzahl an kategorischen Attributen
- $max(n_i)$: Höchste Mindestanforderung des kategorischen Attributs
- $max(n_i)$: Niedrigste Mindestanforderung des kategorischen Attributs

Eine weitere Kennzahl die berechnet wird, ist der Anteilsunterschied. Der Anteilsunterschied, mit dem Buchstaben P abgekürzt, kann einen Wert zwischen 0 und 1 annehmen. Er gibt an, wie groß der Abstand zwischen priviligierten oder bevozugten Gruppen und benachteiligten Gruppen im Bezug auf den Anteil an "1"-Entscheidungen im Durchschnitt ist.

Zur Berechnung der Kennzahl wird für jedes Kategorische Attribut die Differenz zwischen der Ausprägung mit dem höchsten Anteil an "1"- Entscheidungen und der Ausprägung mit dem niedrigsten Anteil an "1"-Entscheidungen berechnet. Diese Differenzen werden aufsummiert und dann anschließend durch die Anzahl an kategorischen Attributen, multipliziert mit 100, geteilt.

Je gleichverteilter die Anteile an "1"-Entscheidungen innerhalb aller kategorischer Attribute sind, desto kleiner wird der Anteilsunterschied. Bei einer kompletten Gleichverteilung nimmt der Anteilsunterschied den Wert 0 an. Analog dazu wird der Anteilsunterschied größer, wenn die Anteile zwischen den Ausprägungen der kategorischen Attribute ungleich verteilt sind.

Angenommen, der Prozess des Artefaktes produziert eine Tabelle mit den kategorischen Spalten "Geschlecht" und "Ethnie". Wenn nun in dieser erzeugten Tabelle exklusiv weiße Männer die "1"-Entscheidung erhalten würden, würde der Anteilsunterschied seinen Maximalwert 1 annehmen.

Der Anteilsunterschied sieht als Formel wie folgt aus:

$$P = \frac{\sum_{i=0}^{m} max(A(x_i)) - min(A(x_i))}{m * 100}$$

Die Variablen in der Formel sind wie folgt zuzuordnen:

- P: Der Anteilsunterschied
- m: Anzahl an kategorischen Attributen
- $A(x)$: Anteil an "1"-Entscheidungen einer Ausprägung x
- $max(A(x_i))$: Höchster "1"-Entscheidungsanteil des kategorischen Attributs
- $max(A(x_i))$: Niedrigster "1"-Entscheidungsanteil des kategorischen Attributs

Da der Anteilsunterschied bei stark verzerrten und diskriminierenden Vorhersagen beziehungsweise vom Artefakt erzeugten Tabellen gegen 1 und bei fairen und gleichbehandelnden gegen 0 geht, spiegelt diese Kennzahl praktisch gesehen die Gerechtigkeit der Vorhersage wider.

4.4 Zusammenfassung

Der ganze Ablauf des Artefakts, von Input zu Output, und die Prozesse des Artefaktes, die während einer Durchführung ablaufen, werden nun, weniger abstrakt als noch in Abbildung 1, in Abbildung 4 visuell zusammengefasst.

Das Artefakt erhält zwei Objekte als Input: einen Datensatz, der den Einschränkungen entspricht und ein Kontextobjekt, welches Metainformationen über die Daten des Datensatzes bereitstellt, wie beispielsweise welche Attribute des Datensatzes numerisch und welche Attribute kategorisch sind.

Der Datensatz wird in einen Test- und in einen Trainingsdatensatz aufgeteilt, mit welchem ein Entscheidungsbaum trainiert wird.

Das Kontextobjekt und der Input-Datensatz werden verwendet, um einen neuen Datensatz zu erzeugen. Dieser Datensatz wird erzeugt, indem das Kreuzprodukt aller Ausprägungen der kategorischen Attribute und die Intervalle der numerischen Attribute gebildet wird. In diesem Datensatz sind nun alle Einträge enthalten, die es basierend auf den Grenzen des Input-Datensatzes geben kann. Das Attribut mit Entscheidungskontext, wie es in den Einschränkungen beschrieben ist, ist kein Attribut des neuen Datensatzes und wurde auch bei der Erzeugung nicht mit einbezogen.

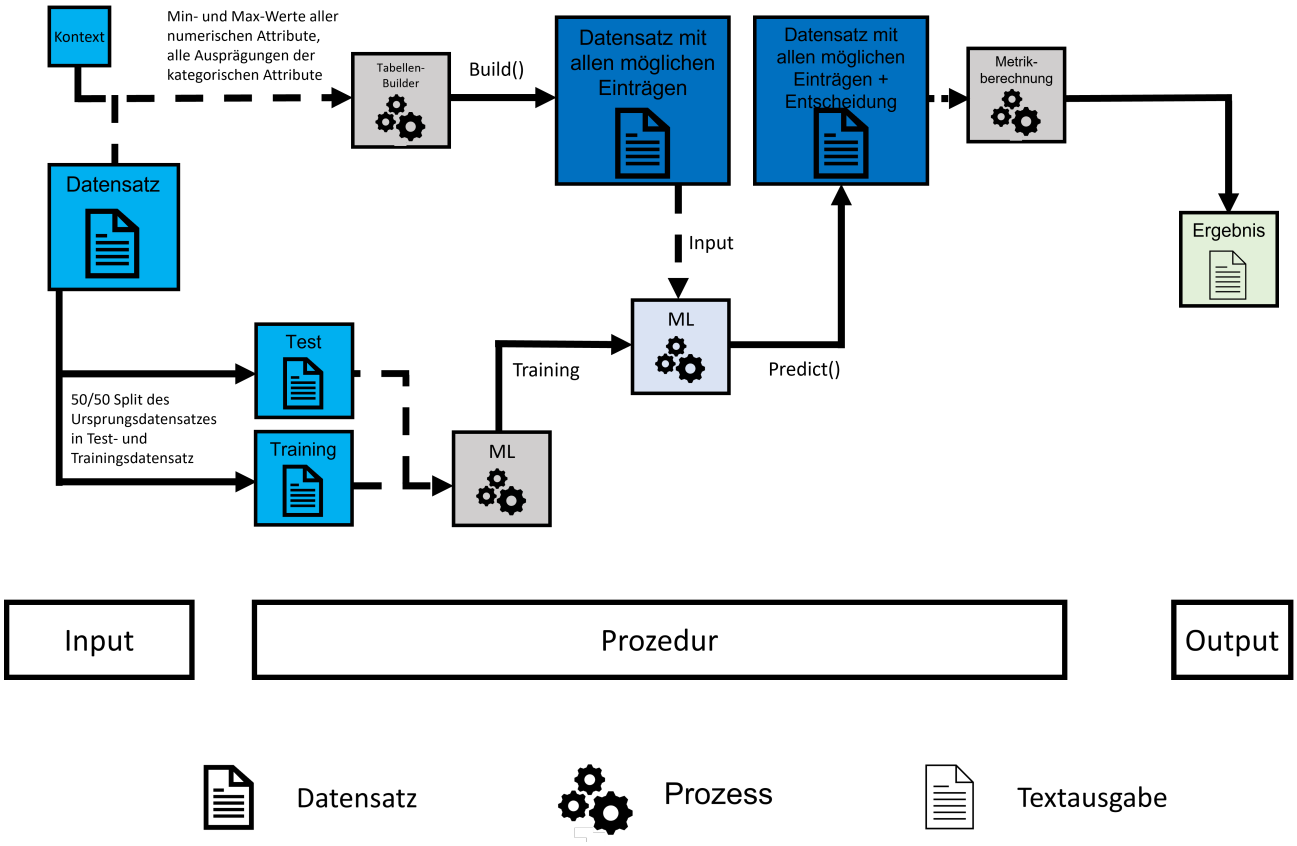

Abbildung 4 Gesamte Ablauf des Artefakts

Für jeden Eintrag im neu erzeugten Datensatz entscheidet der vorher trainierte Entscheidungsbaum, welchen Wert das vorherzusagende Attribut annimmt. Dieser Wert wird für jeden Dateneintrag in einem neu angelegten Attribut gespeichert.

Der um eine weitere Spalte erweiterte erzeugte Datensatz wird nun analysiert. Es werden zu jeder Ausprägung aller kategorischen Attribute die Mindestanforderungen für "1"- beziehungsweise "Ja"-Entscheidung und der Anteil an eben diesen Entscheidungen bestimmt. Darauffolgend werden die beiden Kennzahlen "Anforderungsabstand" und "Anteilsunterschied" berechnet.

All diese erhobenen Analysen und berechneten Kennzahlen werden abschließend in Form einer Textausgabe als menschenlesbarer Output herausgegeben.

5 Evaluation

Für die Evaluation des Artefaktes wurden sechs Datensätze verwendet, über die bereits Metainformationen bezüglich ihrer Gerechtingkeit beziehungsweise Ungerechtigkeit vorliegen. Jeder dieser Datensätze wurde mittels des Artefaktes analysiert und die vom Artefakt erhaltenen Ergebnisse mit den bekannten Metainformationen abgeglichen.

Zwei der sechs Datensätze sind aus dem GitHub Repository der PiML (Python Interpretable Machine Learning) Toolbox. Die beiden Datensätze heißen "CreditSimuBalanced" und "CreditSimuUnbalanced". Bei beiden Datensätzen handelt es sich um simulierte Daten zur Kreditvergabe. Der "CreditSimuBalanced"-Datensatz ist ausgeglichen, also fair und bevorzugt oder benachteiligt keine Gruppe, während der "CreditSimuUnbalanced"-Datensatz hingegen verzerrt und demnach nicht fair ist. Diese beiden Datensätze wurden verwendet, um die Nützlichket des Artefaktes im Bezug auf reelle Anwendungen des echten Lebens zu testen. Obwohl die Datensätze simuliert sind, haben sie einen durchaus realistischen Umfang mit jeweils zehn Attributen, abzüglich personenbezogener Attribute wie Vorname, Nachname oder Kunden-ID. Der "CreditSimuBalanced"-Datensatz hat einen Umfang von 100.000 Dateneinträgen, der andere umfasst 60.526 Dateneinträge.

Die anderen vier Datensätze wurden von uns selbst erstellt. Bei ihnen handelt es sich um fiktive Daten zur Gehaltserhöhungsempfehlung in einem Unternehmen. Die vier Datensätze wurden mit verschiedenen Verzerrungen und Diskriminierungen versehen. Einer der Datensätze enthält absolut keine Verzerrungen und Ungerechtigkeiten, der nächste enthält einen leichten Bias gegebüber einer Personengruppe, der dritte Datensatz enthält mittelstarke Verzerrungen, auch gegen mehrere Gruppen, und der letzte von uns erstellte Datensatz ist sehr ungerecht und bevorzugt eine Personengruppe sehr viel stärker als die anderen. Diese Datensätze wurden verwendet, um die Präzision des Artefaktes bestimmen zu können, da wir exakt wissen, welche Verzerrungen in den Datensätzen enthalten sind und wie sie sich äußern. Alle vier Datensätze sind gleich aufgebaut und weniger umfangreich als die beiden PiML Datensätze. Sie verfügen alle über 5 für die Analyse relevante Attribute und umfassen jeweils 84 Dateneinträge.

Die Datensätze verfügen über die beiden kategorischen Attribute "Geschlecht" und "Ethnie" mit den Ausprägungen $A_G = [m,w]$ und $A_E = [weiss, afroamerikanisch, asiatisch]$ sowie zwei numerische Attribute "Verkäufe" und "Monate beschäftigt" mit den Werteintervallen $I_V = [9; 25]$ und $I_M = [3; 36]$. Für die Anforderungsberechnung wurder als Vergleichswert die Anzahl an Verkäufen durch die Anzahl an Monaten beschäftigt genommen. Das letzte Attribut ist "Gehaltserhöhung" mit den Ausprägungen $A_G =$

$[n,j]$, dieses Attribut ist das Vorherzusagende. Die Zahlenwerte im folgenden Teil sind alle gerundet.

In dem selbsterstellten Datensatz ohne Verzerungen werden alle Personengruppen exakt gleich behandelt, die Dateneinträge sind hier auch praktisch identisch zwischen den Gruppen, mit Ausnahme der kategorischen Attribute. Das Artefakt spiegelt diesen Umstand der Verzerrungslosigkeit auch wider, da sowohl der Anforderungsabstand als auch der Anteilsunterschied 0 beträgt, und die anderen erhobenen Werte für jede Gruppe identisch sind.

Bei dem Datensatz, in den leichte Verzerrungen eingebaut wurden, wurden Männer Frauen gegenüber leicht bevorzugt, sodass sie etwas weniger leisten müssen. Diese Bevorzugung hat sich über alle Gruppen durchgezogen, also wird jeder Mann, unabhängig von seiner Ethnie, gleich bevorzugt. Auch hier spiegelt das Artefakt diesen Umsatnd wider: Zwischen den Ausprägungen der Ethnie sind die "Ja"-Entscheidungen gleichverteilt, während bei den Geschlechtern die Verzerrung erkenntlich wird. 53% der für eine Gehaltserhöhung vorgeschlagenen Personen sind Männer, 47% sind Frauen. Der Anforderungsabstand beträgt 0,021 und der Anteilsunterschied 0,035. Der "Ja"-Entscheidungsunterschied zwischen den Geschlechtern kommt dem des Ursprungsdatensatzes auch sehr nahe, da gehen 56% aller "Ja"-Entscheidungen an Männer und 44% an Frauen.

Der Datensatz mit mittelstarken Verzerrungen bevorzugt Männer genauso stark wie der Datensatz mit leichten Verzerrungen. In diesem Datensatz werden nun zusätzlich noch asiatische Personen benachteiligt. Das Artefakt hat diese Verzerrungen erfasst: Weiße und afroamerikanische Personen machen zusammen 76% der "Ja"-Entscheidungen aus und sind beide jeweils mit 38% vertreten. Asiatische Personen haben nur einen 24%-Anteil an den "Ja"-Entscheidungen. Während eine nicht-asiatische Person hier 0,42 Verkäufe pro Monat für eine "Ja"-Entscheidung braucht, benötigt eine asiatische Person 0,48 Verkäufe pro Monat. Der Anteil der Männer an den "Ja"-Entscheidungen ist im Vergleich zum vorherigen Durchlauf um 3% gestiegen, demnach ist der Anteil der Frauen um 3% gesunken. Beide Geschlechter haben allerdings dieselbe Mindestanforderung für eine "Ja"-Entscheidung. Der Anforderungsabstand beträgt 0,034 und der Anteilsunterschied 0,15. Im Ursprungsdatensatz haben die Männer einen 57%-Anteil an allen "Ja"-Entscheidungen und die Frauen einen 43%-Anteil, was in beiden Fällen eine Abweichung von 1% vom Ergebnis des Artefaktes ist. Die Anteile zwischen den Ethnien stimmen exakt mit denen überein, die das Artefakt ausgegeben hat.

Im letzten von uns erstellten Datensatz sind sehr starkte Verzerrungen eingebaut. Jeder weiße Mann wird hier unabhängig von seiner Leistung für eine Gehaltserhöhung vorgeschlagen, Frauen nie. Männer anderer Ethnie können für eine Gehaltserhöhung vorgeschlagen werden, das passiert jedoch nicht wirklich häufig. Asiatische Männer werden afroamerikanischen Männern gegenüber ganz leicht vorgezogen. Das Artefakt hat ausgegeben, dass weiße Personen einen "Ja"-Entscheidungsanteil von 59% haben, asiatische Personen einen Anteil von 27% und afroamerikanische Personen einen Anteil von 14%. Die Mindestanforderung beträgt bei allen Gruppen 0,25. Nur Männer haben eine "Ja"-Entscheidung bekommen, auch hier ist die Mindestanforderung 0,25. Der Anforderungsabstand beträgt 0 und der Anteilsunterschied 0,73. Im Ursprungsdatensatz haben die weißen Männer einen 63.64%-Anteil an den "Ja"-Entscheidugen, asiatische Männer einen 22,73%-Anteil und afroamerikanische Männer einen 13,64%-Anteil.

Obwohl anfangs von uns vermutet, korreliert der Anforderungsabstand nicht, oder zumindest nicht immer, mit dem Anteilsunterschied. Beim Durchlauf mit dem Datensatz mit starker Verzerrung beträgt der Anforderungsabstand 0, was erstmal widersprüchlch klingt, allerdings lässt sich das durch Rechnungen zeigen. Der Vergleichswert für den Anforderungsabstand wurde wie folgt gewählt: $x = \frac{Verkaeufe}{Monatebeschaeftigt}$. Da weiße Männer unabhängig von ihrer Leistung für eine Gehaltserhöhung vorgeschlagen werden, ist die Mindestvorraussetzung der kleinste Wert des Intervalls I_V geteilt durch den größten Wert des Intervalls I_M. Setzt man dies nun in die Gleichung ein erhält man $x = \frac{9}{36} = 0{,}25$. Nach Betrachtung der vom Artefakt erzeugten Tabelle wurde beobachtet, dass diese Werte auch für die anderen Ausprägungen der Ethnie zu einer "Ja"-Entscheidung führt. Demnach ist dies auch der Mindestwert für die anderen Ethnien. Bei der Berechnung von B steht im Zähler nun 0, da $(0{,}25 - 0{,}25) + (0{,}25 - NaN) = 0$. Der Anforderungsabstand hat sich demnach als eine schwierige Kennzahl herausgestellt, da seine Präzision und sein Nutzen sehr stark vom Kontext der Daten abhängig ist. Gäbe es beispielsweise nur ein einziges numerisches Attribut, von dem die Entscheidung abhängt oder würde die Berechnung eines Vergleichswertes über die Summe aller numerischen Attribute erfolgen, so würde der Anforderungsabstand ein sinnvolles Ergebnis liefern. Der Anforderungsabstand funktioniert so lange, wie der Vergleichswert über eine eindeutige Funktion ohne Ausnahmen definiert ist. Ansonsten ist diese Kennzahl nicht sehr aussagekräftig.

Für die PiML Datensätze terminiert das Artefakt nicht, nach mehreren Stunden wurde die Anwendung beendet. Der vom Artefakt neu erzeugte Datensatz war zu diesem Zeitpunkt ungefähr 50 GB groß. Die numerischen Attribute wurden beim zweiten Versuch als kategorische Attribute gekennzeichnet, allerdings terminierte das Artefakt weiterhin nicht. Aus den Datensätzen wurden darauffolgend kleinere Samples

extrahiert, um Durchläufe mit diesen zu starten. Der Entscheidungsbaum wurde mit dem gesamten Datensatz trainiert.

Zuerst wurde der "CreditSimuBalanced"-Datensatz betrachtet. Bei einer Sample-Größe von 5 nahm der erzeugte Datensatz einen Umfang von 500 bis 20.000 Dateneinträgen an. Zumindest sind das die beobachteten Ausmaße. Unter den beiden Ethnien, die in dem Sample vorkamen, waren die Anteile an "Ja"-Entscheidungen im Schnitt sehr gleichverteilt, das Gleiche gilt auch für die beiden Geschlechtergruppen des Samples. Der Anteilsunterschied ging über die Versuche hinweg gegen 0,0. Für die Sample-Größen von 6 bis 10 wurden Artefaktdurchläufe durchgeführt, die Ergebnisse waren in allen Fällen ziemlich ähnlich und meist auch gleich mit einem Anteilsunterschied um 0,0 herum. Um 0,0 herum beinhaltet übrigens Werte wie 0,03, 0,0000000423 usw., größere Abweichungen wurden über die Versuche hinweg nicht beobachtet. Nach Samplegröße 10 wurden keine Artefaktdurchläufe mehr gestartet, da die Laufzeit und Tabellengröße ab dieser Samplegröße zu groß wurden. Der Durchlauf mit Samplegröße 10 hat eine Tabelle mit 479.999 Einträgen erzeugt.

Bei den Durchläufen für den "CreditSimuUnbalanced"-Datensatz kamen ziemlich genau die gleichen Ergebnisse heraus wie beim "CreditSimuBalanced"-Datensatz, obwohl gegentelige Ergebnisse erwartet wurden. Daraus lässt sich schließen, dass die Sample-Größen nicht ausreichend groß genug waren, um akkurate Aussagen treffen zu können.

Die Durchlaufversuche mit den großen Datensätzen haben gezeigt, dass das Artefakt für große Datensätze praktisch nicht funktioniert. Dies liegt an der vom Artefakt durch das Kreuzprodukt erzeugte Tabelle. Diese Tabelle wächst rapide an, auch bei nur kleinen Änderungen im Datensatz, wie die folgenden Beispiele zeigen:

- Für den selbst erstellten Datensatz wird das Kreuzprodukt $(m,w) * (weiss, afroamerikanisch, asiatisch) * (9,...,25) * (3,...,36)$ berechnet. Die Anzahl an Einträgen, die dieses Kreuzprodukt dann enthält beträgt $2 * 3 * 16 * 33 = 3.168$.
  Der Ursprungsdatensatz hat 84 Einträge.

- Dem Attribut "Geschlecht" wird nun noch die Ausprägung "d" für "divers" hinzugefügt, das Kreuzprodukt steigt dadurch nun an auf $3 * 3 * 16 * 33 = 4.752$ Einträge.

- Der Datensatz erhält nun noch ein zusätzliches Attribut mit drei Ausprägungen. Das Kreuzprodukt steigt dadurch nun an auf $3 * 3 * 3 * 16 * 33 = 14.256$ Einträge.

- Der "CreditSimuBalanced"-Datensatz enthält zehn für die Rechnung relevante Attribute, die Berechnung des Kreuzproduktes muss mit $(104.698,...,1.822.208) * (2,...,16.839) * (0,...,12640) * (0,...,5) * (0,...,11) * (0,1) * (0,1) * (0,1) * (0,1)$ berechnet werden. Dadurch ergibt sich die Rechnung $1.717.510 * 16.839 * 12641 * 6 * 12 * 2 * 2 * 2 * 2 = 4{,}21162295e^{17}$. So viele Einträge würden für diesen Datensatz erstellt werden. Angenommen, jeder Dateneintrag der erzeugten CSV hätte durchschnittlich 30 Zeichen, dann wäre die Dateigröße $30B * 4.21162295e^{17} = 12.634.868.795{,}9GB = 12.634{,}8687959PB$

Die erzeugte Tabelle kann sehr schnell sehr große Ausmaße annehmen, besonders starken Einfluss hat hierbei die Anzahl an Attributen. Aufgrund dieses starken Wachstums ist das Artefakt in der Regel nicht sinnvoll auf reale Datensätze zu verwenden, da diese meist mehr als nur vier oder fünf Attribute haben.

6 Diskussion

Die in der Evaluation beschriebenen Versuche und deren Ergebnisse zeigen Vieles über das entwickelte Artefakt auf.

Die Ergebnisse des Artefaktes waren bisher sehr präzise, was die Durchläufe mit den selbsterzeugten Datensätzen gezeigt haben, besonders im Bezug auf die Anteilsverteilungen. Demnach ist auch der Anteilsunterschied eine gute Kennzahl, da er auf sehr präzisen Ergebnissen aufbaut. Die andere von uns erhobene Kennzahl, der Anforderungsabstand, ist stark vom Vergleichswert und den Daten abhängig. Kann man die Daten mittels einer ganz genauen Funktion normalisieren und vergleichen, so geben der Anforderungsabstand und die berechneten Mindestvorraussetzungen sinnvolle Ergebnisse. Ist dies allerdings nicht möglich, sind beide Werte nicht sehr aussagekräftig. Die beobachtete Präzision des Artefakts, vor allem in Bezug auf den Anteilsunterschied und alles, was dort dazugehört, zeigt, dass sich der Bias eines Datensatzes, auch wenn er nur sehr leicht ist, auch schon bei kleineren Datenmengen auf das ML-Modell übertragen lässt. Demnach lassen sich dadurch Rückschlüsse auf den Ursprungsdatensatz ziehen. Es ist zwar nicht möglich zu sagen, warum oder wo genau Verzerrungen vorliegen, aber es lässt sich sagen, welche Verzerrungen im Entscheidungsbaum enthalten sind, wenn dieser Datensatz zum trainieren eines Entscheidungsbaumes vewendet wird.

Es scheint sehr wahrscheinlich, dass bei Datensätzen, in denen manche Personengruppen unterrepräsentiert sind, aber fair behandelt werden, auch eine Verzerrung vom Artefakt entdeckt wird, obwohl keine enthalten ist. Demnach würden in dem Fall fälschlicherweise Verzerrungen erkannt werden. Sollte das passieren, zeigt das in erster Linie, dass der Datensatz nicht geeignet ist, um einen Entscheidungsbaum damit zu trainieren, da der Entscheidungsbaum aufgrund der Unterrepräsentation Verzerrungen gebildet hat. Diese Vermutung ist bei der Auswertung der durchgeführten Durchläufe entstanden, aufgrund von zeitlichen Einschränkungen konnte diese nicht mit weitergehenden Versuchen geprüft werden.

Die Versuche mit den PiML Datensätzen haben gezeigt, dass das Artefakt für große und aufwendige Datensätze nicht terminiert, wodurch es stark an Nutzen einbüßt. Das starke Anwachsen des erzeugten Datensatz ist kritisch, wodurch es nicht möglich ist, einen realen Datensatz zu nehmen und durch das Artefakt durchlaufen zu lassen. Um dennoch solche Datensätze verwenden zu können, müsste man diese reduzieren und auf einen kleineren Datensatz abbilden, indem man Attribute zusammenfasst oder weglässt und nur Samples von den Dateneinträgen nimmt, aber auch dann könnte das

Artefakt sehr lange zum terminieren brauchen. Demnach ist eine Einschränkung des Artefakts die Größe und die Anzahl an Attributen des Datensatzes.

Aus theoretischer Sicht funktioniert die Methode und sie beantwortet auch die Forschungsfrage mit einer möglichen Umsetzungsmöglichkeit. Verzerrungen, die in Datensätzen enthalten waren, konnten mit einer hohen Präzision bestimmt werden. Das haben die Durchläufe mit den selbsterzeugten Datensätzen gezeigt.

Aus praktischer Sicht ist dieser Ansatz so, wie er gerade ist, in den meisten Fällen unbrauchbar. Bei großen und komplizierteren Datenmengen terminiert die Methode nicht und erzeugt eine so große Datei, dass eine herkömmliche Festplatte nicht ausreicht, um die Methode laufen zu lassen.
Für diese Methode müsste man weitere Arbeit in den Schritt der Tabellenerzeugung stecken. Der Ansatz, alle möglichen Möglichkeiten abzubilden, führt zwar zu einer hohen Präzision, aber ist zu zeit- und speicherintensiv. Ein möglicher Alternativansatz für diesen Schritt wäre eine randomisierte Umsetzung. Bei dieser könnte man ein Limit an Dateneinträgen in der erzeugten Tabelle einstellen, wenn das Kreuzprodukt dieses Limit überschreiten würde, wird nicht das Kreuzprodukt gebildet, stattdessen wird der erzeugte Datensatz mit zufälligen Dateneinträgen befüllt, bis das Limit erreicht ist. Wenn das Kreuzprodukt das Limit nicht überschreitet, wird das Kreuzprodukt verwendet. Dieser Ansatz wäre zwar weniger präzise als der aktuelle, allerdings könnte er tatsächlich terminieren.
Auch müssten mehr Versuche mit unterschiedlicheren Datensätzen durchgeführt werden, um die Ergebnisse weiter bestätigen zu können.

7 Fazit

Verzerrungen und Diskriminierung in Datensätzen sind ein Problem, auch im Machine Learning Bereich, das war schon vor dieser Arbeit bekannt. Die Durchläufe mit den selbst erzeugten Datensätzen haben noch einmal gezeigt, wie sehr sich auch leichte und vor allem leicht übersehbare Verzerrungen auf Machine Learning Modelle übertragen, auch wenn die Datenmenge klein ist und auch nur eine Teilmenge des Datensatzes zum Trainieren verwendet wird. Die von uns konzipierte Methode zur Überprüfung von Verzerrungen und Bewertung der Stärke der Verzerrunen hat sich als erstaunlich präzise herausgestellt. Aufgrund ihrer Gründlichkeit kann die Methode herausfinden, welche Personengruppen in einem Datensatz privilegiert sind und welche Personengruppen benachteiligt werden, falls solche Verzerrungen vorhanden sind. Direkte Aussagen über den Input-Datensatz lassen sich nicht treffen, wenn man die Ergebnisse der Methode betrachtet, da die Methode im Analyseschritt nur den Datensatz, der von der Methode selbst erstellt wird, betrachtet. Um von den Ergebnissen direkt auf den Datensatz schließen zu können, müssten sehr viel mehr Testfälle und Szenarien durchgetestet werden, um diesen Folgeschluss erlauben zu können. Ob ein Datensatz nun eine Personengruppe benachteiligt, kann nicht direkt gesagt werden, nachdem die Methode durchgeführt wurde, aber es kann die Aussage getroffen werden, dass wenn ein Entscheidungsbaum mit diesem (Teil-)Datensatz trainiert wird, dessen Modell die von der Methode ausgegebenen Verzerrungen hat. Dennoch konnten in den Testdurchläufen mit den selbsterstellten Datensätzen alle Verzerrungen, die in den Daten vorlagen, in der Ergebnisausgabe der Methode wiedergefunden werden.

Die Methode ist in der Praxis so, wie sie bisher aufgebaut ist, nicht funktionstüchtig, da sie für größere und aufwendigere Datensätze nicht terminiert und eine Datei erzeugt, die so groß werden kann, dass praktisch kein Computer, der diese Methode ausführt, diese Datei speichern kann.

Abschließend kann man in Bezug auf die Forschungsfrage sagen, dass eine funktionierende Methode entworfen wurde, die Verzerrungen und potentielle Diskriminierungen in tabellarischen Datensätzen mittels Verzerrungsübertragung auf eine Machine Learning Modell erkennen und bewerten kann. Sie ist nur in der Praxis aufgund ihrer Defizite nicht anwendbar.

Literaturverzeichnis

Agarwal, S., Muku, S., Anand, S., & Arora, C. (2022). Does data repair lead to fair models? curating contextually fair data to reduce model bias. *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*, 3898–3907. https://doi.org/10.1109/WACV51458.2022.00395

Beattie, H., Watkins, L., Robinson, W., Rubin, A., & Watkins, S. (2022). Measuring and mitigating bias in ai-chatbots. *Proceeding - 2022 IEEE International Conference on Assured Autonomy, ICAA 2022*, 117–123. https://doi.org/10.1109/ICAA52185.2022.00023

Brotcke, L. (2022). Time to assess bias in machine learning models for credit decisions. *Journal of Risk and Financial Management*, *15*. https://doi.org/10.3390/jrfm15040165

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255

Friedler, S. A., Choudhary, S., Scheidegger, C., Hamilton, E. P., Venkatasubramanian, S., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 329–338. https://doi.org/10.1145/3287560.3287589

Gregor, S., & Hevner, A. R. (2013). *Positioning and presenting design science research for maximum impact* (2) [DSR Study].

He, Y., Burghardt, K., & Lerman, K. (2020). A geometric solution to fair representations. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 279–285. https://doi.org/10.1145/3375627.3375864

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*, 1–33. https://doi.org/10.1007/s10115-011-0463-8

Mazilu, L., Paton, N., Konstantinou, N., & Fernandes, A. (2020). Fairness in data wrangling. *Proceedings - 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science, IRI 2020*, 341–348. https://doi.org/10.1109/IRI49571.2020.00056

Mehrabi, N., Gupta, U., Morstatter, F., Steeg, G. V., & Galstyan, A. (2022). Attributing fair decisions with attention interventions. *TrustNLP 2022 - 2nd Workshop on Trustworthy Natural Language Processing, Proceedings of the Workshop*, 12–25.

Robert, L., Pierce, C., Marquis, L., Kim, S., & Alahmad, R. (2020). Designing fair ai for managing employees in organizations: A review, critique, and design agenda. *Human-Computer Interaction*, *35*, 545–575. https://doi.org/10.1080/07370024.2020.1735391

Roselli, D., Matthews, J., & Talagala, N. (2019). Managing bias in ai. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, 539–544. https://doi.org/10.1145/3308560.3317590

Abschließende Erklärung

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Nürnberg, den 10. Februar 2023

Rene Jokiel, Philippe Huber