FIfF-Studienpreis 2017
1. Preis

Tobias D. Krafft

Qualitätsmaße algorithmischer Entscheidungssysteme in der Kriminalprognostik

In meiner Masterarbeit vom Juni 2017 mit dem Titel Qualitätsmaße binärer Klassifikatoren im Bereich kriminalprognostischer Instrumente der vierten Generation *(Krafft 2017) setze ich mich kritisch mit der Integration von algorithmischen Entscheidungssystemen (ADM) in der Kriminalprognostik auseinander.*

1. Einleitung

In der heutigen digitalen Gesellschaft unterstützen algorithmische Entscheidungssysteme, ADM (*algorithmic decision making*) genannt, vgl. Zweig (2016), zunehmend die Entscheidungsfindung in den verschiedensten Bereichen (Lischka und Klingel 2017), auch in der Justiz. Verlässt sich nun eine Gesellschaft in einem derart relevanten Bereich wie der Justiz in der Urteilsfindung auf algorithmenbasierte Risikoprognosen, so muss deren Evaluation ein zentrales Anliegen sein.

Die Arbeit verfolgt daher das Ziel, den aktuellen Stand um die Beurteilung von ADM-Prozessen im kriminalprognostischen Bereich zu erfassen und kritisch zu hinterfragen. Wegweisend für mein Vorgehen sind die im *ADM-Manifest* (AlgorithmWatch 2016) der 2016 gegründeten Bürgerinitiative *AlgorithmWatch* differenzierten Aspekte als Bestandteil der algorithmischen Entscheidungsfindung.

Während in den USA schon seit längerem ADM-gesteuerte Kriminalprognosen eingeführt sind, nutzen deutsche Gerichte aufgrund des *Individualisierungsgebots* (Maschke 2008) in der deutschen Rechtsprechung individuelle kriminalprognostische Gutachten, die durch entsprechende Sachverständige erstellt und bei gewissen strafrechtlichen Entscheidungen miteinbezogen werden müssen[1]. Es ist jedoch zu vermuten, dass sich dies im Hinblick auf die Aufmerksamkeitsökonomie sowohl der Richter als auch der Sachverständigen in absehbarer Zeit ändern könnte, sollte die deutsche Justiz mit Argumenten der Effizienz und angeblicher Objektivität vom Einsatz ADM-gestützter Prognoseinstrumente überzeugt werden. Insofern war es ein zentrales Anliegen der Arbeit, eine weitere Diskussion über Chancen und Risiken beim Einsatz algorithmenbasierter Risikoprognosen auf wissenschaftlicher Grundlage zu ermöglichen; siehe auch Zweig et al. (2018).

Dazu wird im Folgenden das üblicherweise verwendete Qualitätsmaß zur Bewertung von Klassifikatoren, die sogenannte *Area under the Receiver Operating Characteristic (AUC)* kritisch hinterfragt und dem *Positive Predictive Value among the first k* (PPV_k) als weitere Bewertungsalternative, die zudem als Evaluationsmöglichkeit den richterlichen Entscheidungsprozess realistischer abbildet, gegenübergestellt. Die Überprüfung, ob es Klassifikatoren gibt, für die diese Qualitätsmaße zu unterschiedlichen Ergebnissen kommen können, kam zu frappierenden Ergebnissen, denn es ergaben sich mögliche Abweichungen von bis zu 0,75. Die Diskrepanz zwischen den beiden Qualitätsmaßen konnte an einem realen Datensatz aus den USA für das sogenannte COMPAS-Tool, ein kriminalprognostisches Instrument der vierten Generation, nachgewiesen werden.

Im abschließenden Fazit werden mögliche gesellschaftliche Folgen aufgezeigt und versucht, Handlungsempfehlungen und Lösungsvorschläge zu umreißen.

1.1 Die Integration von ADM-gestützten Risikoprognosen im US-amerikanischen Justizwesen

Das Justizsystem der USA ist im Begriff, zu kollabieren: Als Folge der etablierten Praxis, drakonische Strafen zur Abschreckung zu nutzen, sitzen derzeit in US-amerikanischen Gefängnissen knapp 2,15 Millionen Inhaftierte (Statista 2017) ein. Die hohe Zahl von Gefängnisinsassen beschert den USA explodierende Kosten im Strafvollzug, sodass die Suche nach einer effizienten Kostenreduzierung ein zentrales Anliegen der US-amerikanischen Justiz ist.

Die Erkenntnis, kostenintensive Haftstrafen auf Bewährung auszusetzen, ließ die Kriminalprognose zunehmend in den Fokus juristischer Überlegungen rücken, und man forderte bei der Urteilsfindung eine genauere Einschätzung des Kriminalitätsrisikos (Chettiar und Gupta 2011). Daher benutzt die Justiz aller US-Bundesstaaten seit Jahren Tools zur Risikobewertung in verschiedenen Bereichen der Rechtsprechung (EPIC 2017), wobei von einigen Behörden, wie zum Beispiel in Florida, hauptsächlich das von der US-Firma *Northpointe* Ende der 1990er-Jahre entwickelte *COMPAS Assessment Tool (Correctional Offender Management Profiling for Alternative Sanctions)* zum Einsatz kommt (Northpointe 2012b).

2. Problematik und Instrumente der Kriminalprognose

In früheren Zeiten oblag die Rechtsprechung allein dem Richter und etwaigen Beratern. Diese nur auf der subjektiven Entscheidung des Richters basierende Urteilsfindung war extrem anfällig für Fehlurteile, auch hinsichtlich der Höhe des verhängten Strafmaßes, sodass man seit der Aufklärung (Cesare 1764) versuchte, objektive Kriterien für die Urteilsfindung zu entwickeln. Doch trotz aller Bemühungen sieht sich auch heute noch jede Kriminalprognose mit dem Problem konfrontiert, dass eine hundertprozentige Vorhersage nicht garantiert werden kann. Das menschliche Verhalten resultiert eben nicht nur aus individuellen Persönlichkeitsmerkmalen, sondern wird dazu durch verschiedenste situative Faktoren beeinflusst (Danziger et al. 2011), die aufgrund ihrer Variabilität nur vage abschätzbar sind (Bliesener et al. 2014, S. 425f.).

Der Baxstrom-Fall aus dem Jahr 1966 sensibilisierte weltweit auch die Öffentlichkeit für diese Problematik. Bei dem unbeabsichtigten Experiment mussten aus formaljuristischen Gründen der Gewalttäter Johnnie Baxstrom sowie 967 weitere, als gefährlich eingeschätzte Straftäter im Bundesstaat New York freigelassen werden. Nach insgesamt vier Jahren in Freiheit waren aber lediglich 14,2 % der als gefährlich eingestuften Täter erneut straffällig geworden, darunter nur 2,5 % wegen schwerer

Gewaltstraftaten (Obergfell-Fuchs 2011, S. 17; US Supreme Court 1966).

2.1 Instrumente der Kriminalprognostik

Im Bestreben, die Kriminalprognose auf eine überprüfbare Basis zu stellen, wurden eine Vielzahl von Prognoseinstrumenten entwickelt (Guy 2008), sodass hier der in der Fachliteratur üblichen Einordnung in ‚Generationen' grob gefolgt wird (Döbele 2014, S. 20–26; Rettenberger und Franqué 2013, S. 21f.).

Anfang des 20. Jahrhunderts wurden erste Kriterienkataloge entwickelt, sogenannte Prognosetafeln, mit denen man potenzielle Straftäter zu identifizieren hoffte (Nedopil und Gross 2005, S. 43f.) und die den Grundstein zur statistischen Kriminalprognose legten. Sie basierten auf rückfallrelevanten Faktoren, die aus Akten entlassener Straftäter extrahiert wurden. Diese statischen Merkmale wurden von der zweiten Generation durch personen- und tatbezogene Faktoren ergänzt, dennoch blieben etwaige Wandlungen der Täterpersönlichkeit weiterhin unberücksichtigt, sodass der Straftäter hier „zum Gefangenen seiner Biographie" (Dittmann 2003) wurde. Folgerichtig führte die dritte Generation zu Instrumenten, welche die Datenbasis um dynamische Faktoren, wie „persönliche Einstellung der Straftäter, […], soziale Bindungen" usw. (Döbele 2014) erweiterten.

Die vierte Generation repräsentiert den aktuellsten Stand der Kriminalprognostik und sieht ein breites Band an Einsatzmöglichkeiten für verschiedenste Prognosebereiche vor. Zum einen fließen immer mehr variable Aspekte in den Beurteilungsprozess ein, zum anderen beschränken sich die Tools nicht mehr nur auf Risikoprognosen von Verhaltensweisen, sondern bieten Empfehlungen für Therapiepläne an oder werben sogar damit, dass sie Aussagen machen könnten, ob ein Straftäter vor Gericht erscheine oder nicht, vgl. z. B. Northpointe (2012b).

2.2 COMPAS als Instrument der vierten Generation

Ein bekanntes Vorhersageinstrument der vierten Generation ist die *Correctional Offender Management Profile for Alternative Sanctions,* kurz COMPAS genannte Web-Applikation, die vom *Northpointe Institute for Public Management Inc.* als automatisierte Entscheidungsunterstützung zur Bewertung von Straffälligen entwickelt wurde (Northpointe 2012a). Die Firma wirbt mit dem Angebot, mit Hilfe ihres Algorithmus ließe sich auf Basis von 137 Merkmalen eine genaue Prognose der Rückfallwahrscheinlichkeit (*predicted recidivism*) eines Angeklagten erstellen (Brennan et al. 2009b), so dass einige Staaten der USA diese bereits im juristischen Prozess einsetzen, um beispielsweise Richter bei Bewährungsfragen zu beraten.

Wie bei fast allen Instrumenten der vierten Generation ist jedoch auch dieser Algorithmus proprietär und folglich eine ‚Blackbox' (Diakopoulos 2014), sodass die geringe Transparenz der Algorithmen und die daraus resultierende fehlende Einsicht in den Bewertungsprozess auch beim COMPAS-Tool ein großes Problem darstellt. Mögliche Überprüfungsstrategien sind sogenannte Blackbox-Analysen, die eigentlich aus der Softwareentwicklung stammen, vgl. Beizer (1995), und bei denen ohne Kenntnis der inneren Funktionsweise der Algorithmen die tatsächlichen Ergebnisse mit den zu erwartenden überprüft werden.

ProPublica, eine durch Spenden finanzierte US-Rechercheorganisation, hat 2016 eine Studie veröffentlicht, die für das COMPAS-Tool eine dramatische Ungleichbehandlung von Schwarzen und Weißen nachgewiesen hat (Angwin et al. 2016). Die Antwort aus dem Hause Northpointe (Dieterich et al. 2016) begründete die festgestellten Ergebnisse mit der Nutzung eines anderen Fairnesskriteriums und zeigt so den sehr weiten Modellierungsrahmen auf, dem solche Instrumente unterliegen. An dieser Stelle sei noch angemerkt, dass die unter *Northpointe* bekannte und erfolgreiche Firma kurz nach der Debatte aus ungeklärten Gründen ihren Namen in *Equivant* geändert hat.

3. Abweichung zwischen *AUC* und PPV_k

Die Kontroverse von 2016 zeigt die dringende Notwendigkeit, den Bewertungsprozess dieser Algorithmen näher zu beleuchten. Anderenfalls ist keine realistische Einschätzung möglich, ob und wann ein solches ADM zur Beurteilung von Menschen in einem so essenziellen Bereich wie der Justiz eingesetzt werden sollte.

Die in der Justiz Anwendung findenden Instrumente münden in binäre Klassifikatoren[2], da die juristische Fragestellung nur duale Urteile zulässt: Schuldig oder nicht. Das bestehende Repertoire

Tobias D. Krafft

Tobias D. Krafft, M.Sc., ist Doktorand am Lehrstuhl *Algorithm Accountability* von Prof. Katharina A. Zweig an der TU Kaiserslautern. Als Preisträger des Studienpreises 2017 des Forum InformatikerInnen für Frieden und gesellschaftliche Verantwortung reichen seine Forschungsinteressen von der (reinen) Analyse algorithmischer Entscheidungssysteme bis hin zum Diskurs um deren Einsatz im gesellschaftlichen Kontext. Im Rahmen seiner Promotion hat er das Datenspendeprojekt mitentwickelt und einen Teil der Datenanalyse durchgeführt. Er ist einer der Sprecher der Regionalgruppe Kaiserslautern der Gesellschaft für Informatik, die es sich zur Aufgabe gemacht hat, den interdisziplinären Studiengang der Sozioinformatik (TU Kaiserslautern) in die Gesellschaft zu tragen. Zuschriften an *krafft@cs.uni-kl.de*

zur Bewertung solcher Klassifikatoren ist zwar differenziert, jedoch wurden bereits 1977 die ersten Rückfälligkeitsvorhersage-Statistiken mit der *Area under the Receiver Operating Characteristic (AUC)* bewertet (Fergusson et al. 1977). Dieses im maschinellen Lernen häufig verwendete Qualitätsmaß hat sich in der Kriminalprognose als vorherrschend etabliert.

Obwohl es sich hierbei eigentlich um eine Betrachtung der Sensitivität und Falsch-Positiv-Rate handelt (Aggarwal 2015, S. 340f.; Bradley 1997, S. 2; Peterson et al. 1954), lässt sich die *AUC* bei der Bewertung binärer Klassifikatoren zusätzlich als die Wahrscheinlichkeit interpretieren, in einer zufälligen Stichprobe, bestehend aus je einem Element beider Klassen, dem Element der ersten Klasse eine höhere Wahrscheinlichkeit zuzuordnen, zu dieser zu gehören (Hanley und McNeil 1982, S. 2). Hat ein ADM-System also eine *AUC* von 0,72, so kann es, gegeben einen rückfällig werdenden Straftäter und einen, der es nicht wird, mit einer Wahrscheinlichkeit von 72 % korrekt entscheiden, welcher von beiden der rückfällig Werdende ist.

Dennoch ist ihr Ruf als bestes Qualitätsmaß der Kriminalprognose (Barnoski und Drake 2007) insgesamt nicht nachvollziehbar, auch die uneinheitliche Anwendung der *AUC* in verschiedenen Disziplinen erstaunt. So verwendet die Humanmedizin andere Schwellenwerte (Leushuis et al. 2009) als die Risikoprognose. Es ist nicht zu verstehen, warum diese Werte je nach Disziplin differieren (Leushuis et al. 2009; Endrass et al. 2008) und warum bei kriminalprognostischen Instrumenten ein deutlich niedrigerer Wert (0,65–0,75) (Lansing 2012, S. 22) als Garant für eine gute Klassifikation angesehen wird.

Aber selbst, wenn ein Klassifikator eine *AUC* von 1,00 erreicht, könnte er jedem Rückfälligen eine Rückfallwahrscheinlichkeit von 10 % und jedem nicht Rückfälligen eine von 9 % prognostizieren. Obwohl laut *AUC* diese Klassifizierung eine perfekte Trennschärfe aufweist, ließe sich einerseits sehr schwer zwischen den beiden Klassen separieren, andererseits ist eine solch niedrige Rückfallwahrscheinlichkeit in keiner Art hilfreich, wenn die Basisrate höher liegt als die Prognose.

Diese Diskrepanz würde sich zwar auch bei einer Nutzung des PPV_k widerspiegeln, jedoch bildet dieser, wie im Folgenden erläutert wird, den generellen Entscheidungsprozess eines Richters deutlich besser ab.

Der *Positive Predictive Value among the first k (PPV_k)* stellt eine andere, in der Kriminalprognostik allerdings kaum beachtete Möglichkeit dar, einen Klassifikator zu bewerten. Allein die Eigenschaft „ist unter den ersten k", also am höchsten gerankt, ist hier relevant, sodass der PPV_k eine Möglichkeit bietet, den Fokus auf die tatsächliche Anzahl korrekt klassifizierter Objekte im vorderen Bereich der Sortierung zu legen. Da er fast vollständig auf eine Betrachtung der genauen Sortierung der Elemente, vor allem im hinteren Teil, verzichtet, wird er dem richterlichen Entscheidungsprozess zudem deutlich gerechter, denn **auch Richter folgen wahrscheinlich einer inneren Skala,** um für einen Straftäter zu bestimmen, ob dieser rückfällig werden würde oder ab welchem Bereich Straftätern eine Bewährung gewährt wird oder nicht.

Sowohl aufgrund seiner Eignung zur Bewertung eines Klassifikators als auch seiner Nähe zum richterlichen Entscheidungsprozess wurde der PPV_k gewählt, um der *AUC* mathematisch gegenübergestellt zu werden. Das Ergebnis der Analyse zur Klärung, wie weit die beiden Werte für einen gegebenen Klassifikator voneinander abweichen können, ist besorgniserregend: Abweichungen von bis zu 0,75 sind nachweislich möglich.[3]

Die Abweichung dieser beiden Maße ist auch insofern von Relevanz, als Northpointe aktuell nur die hohe *AUC* (Northpointe 2012b) angibt, um ihre vermeintlich guten Klassifikatoren zu bewerben.

4. Überprüfung der Ergebnisse für das COMPAS Assessment Tool

Die aufgezeigten Ergebnisse geben zwar für eine binäre Klassifizierung Aufschlüsse über die Lage der jeweiligen maximal/minimal möglichen PPV_k-Werte bei fixierter *AUC* und vice versa, jedoch können mittels dieser Formeln noch keine Rückschlüsse darüber gezogen werden, wie sich die Verteilung zwischen diesen Werten darstellt. Um die gesellschaftliche Brisanz der möglicherweise fehlenden Korrelation anhand von anwendungsbezogenen Daten zu erörtern, wurden die von ProPublica öffentlich bereitgestellten Datensätze extrahiert und entsprechend analysiert. Da hierbei das momentan in Wisconsin, USA, auf jeder Stufe des Justizprozesses angewandte (Wisconsin Department of Correction 2018) COMPAS Assessment Tool anhand echter Datensätze evaluiert werden kann, lässt sich überprüfen, ob im Anwendungsbezug die Korrelation von *AUC* und PPV_k gewahrt bleibt oder sie deutlich voneinander abweichen. Es muss darauf hingewiesen werden, dass den folgenden Ergebnissen lediglich ein Algorithmus und ein lokal erfasster Datensatz zugrunde liegt.

4.1 Erläuterung zum COMPAS Assessment Tool

Das Tool wurde 1998 als „Breitband"-Bewertung konzipiert und kann anhand verschiedener Fragebögen in 22 verschiedenen Bedürfnis- und Risikobereichen Prognosen über Individuen erstellen (Northpointe 2012a), unter anderem die folgenden:

- *General Recidivism Risk Scale:* Generelles Rückfallrisiko *(GRRS)*
- *Violent Recidivism Risk Scale:* Gewaltbasiertes Rückfallrisiko *(VRRS)*

Die kontinuierliche Vorhersage der einzelnen Skalen wird im Justizwesen auf binäre Entscheidungen abgebildet, z. B. auf Fragen, ob der verurteilte Straftäter auf Bewährung das Gefängnis verlassen darf oder nicht. Somit bietet sich in der Evaluation das Spektrum eines binären Klassifikators an, weshalb Northpointe (2012b) selbst die hohe *AUC* seines Algorithmus lobt und mit der *AUC* in verschiedenen Studien wirbt, die den COMPAS GRRS mit den in Tabelle 1 aufgeführten *AUC*-Werten beurteilen (Northpointe 2012b).

Quelle	Jahr	Stichprobengröße	Betrachtungszeitraum der Rückfälligkeit	AUC
(Brennan et al. 2009a) (Brennan et al. 2009b)	2009	2.328	1 Jahr	0,68
(Farabee et al. 2010)	2010	25.009	2 Jahre	0,70
(Lansing 2012)	2012	11.289	2 Jahre	0,71

Tabelle 1: Auszug der von Northpointe (2012b) veröffentlichten Liste an Evaluationen des COMPAS-Assessment-Tools

4.2 Auswertung der ProPublica-Daten zum COMPAS Assessment Tool

Dem vorliegenden ProPublica-Datensatz konnten für 11.777 Personen alle notwendigen Informationen entnommen werden, um sowohl die *AUC* als auch den erreichten PPV_k zu bestimmen[4].

Wenn angenommen wird, dass ein Richter in seinem Distrikt den vorliegenden Datensatz zu bearbeiten hat, so hätte er bei der zufälligen Verurteilung eines Delinquenten entsprechend der im Datensatz vorherrschenden Basisrate eine Wahrscheinlichkeit von 36 %, dass dieser rückfällig wird. Nutzt er nun das vorgestellte COMPAS-Tool und kann auf Grund seiner Erfahrung mit dem Distrikt die Basisrate korrekt abschätzen, verurteilt also nur die Straftäter, die in den oberen 36 % der Datenpunkte liegen, so erreicht er, der Interpretation des PPV_k folgend, lediglich mit einer Wahrscheinlichkeit von ungefähr 53 % (siehe Abbildung 1) einen wirklich rückfällig Werdenden.

Viel drastischer stellt sich der Unterschied zwischen *AUC* und PPV_k für die *Violent Recidivism Risk Scale (VRRS)* dar. In Abbildung 2 ist durch den grünen Punkt zu erkennen, dass es Northpointe mit diesem Klassifikator deutlich in das untere Drittel des möglichen Wertebereichs für den PPV_k mit vorliegender *AUC* von 0,69 schafft. Es ist jedoch zu beachten, dass der Klassenunterschied mit geringer Basisrate von 8 % deutlich größer ist, sodass es wiederum schwieriger wird, eine Klassifizierung zu schaffen, welche die wenigen Rückfälligen von den vielen nicht Rückfälligen separiert. Ein Richter würde hier durch zufällige Urteile einen tatsächlich gewaltbasiert Rückfälligen lediglich mit einer der Basisrate entsprechenden Wahrscheinlichkeit von 8 % erfassen. Dieser Wert kann zwar durch die Anwendung des COMPAS-Tools um 11,6 Prozentpunkte erhöht werden, jedoch liegt die Trefferwahrscheinlichkeit innerhalb der ersten 1.085 Straffälligen immer noch bei nur 20 % und gibt dem Anwender keine wirklich stichhaltige Prognose, auf der sein Urteil aufgebaut werden könnte.

5. Fazit

Die hohe Nutzungsquote der *AUC* bei binären Klassifikatoren für Rückfälligkeitsvorhersagen erscheint willkürlich und nur durch die Verbreitung beim maschinellen Lernen begründet. Auch die Überprüfung der *AUC* hinsichtlich ihrer Eignung als

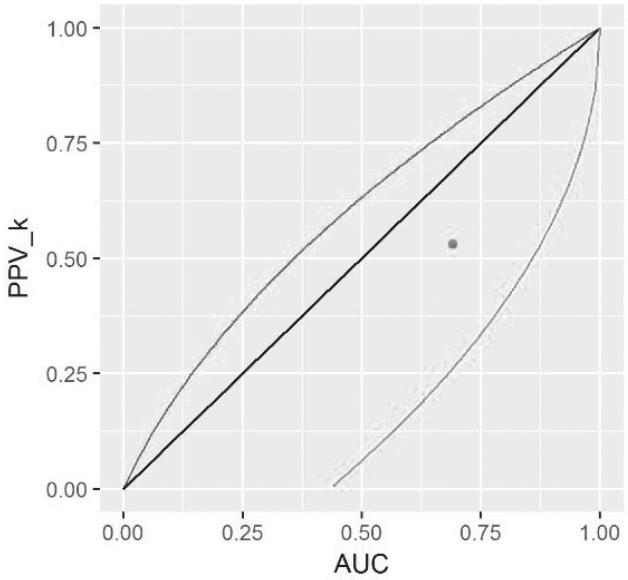

Abbildung 1: Auswertung des COMPAS-Scores (GRRS) auf den ProPublica-Daten (grüner Punkt) bei eingezeichnetem maximalen PPV_k (obere Kurve, blau) und minimalem PPV_k (untere Kurve, rot) für jeweils fixierte AUC auf der x-Achse, Abbildung aus (Krafft 2017).

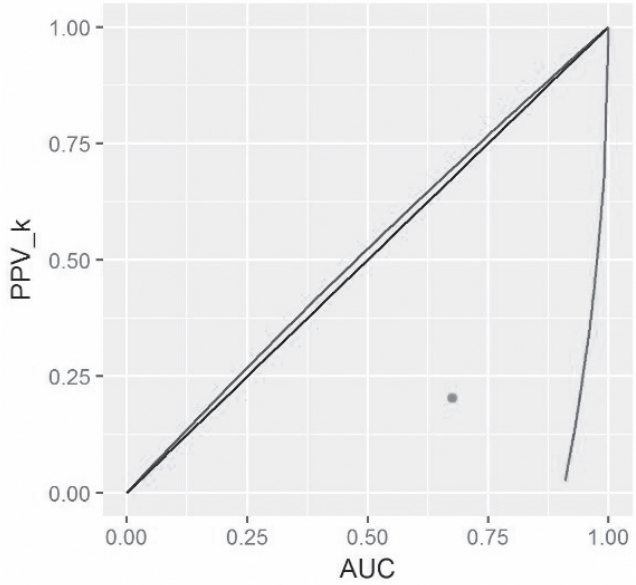

Abbildung 2: Auswertung des COMPAS-Scores (VRRS) auf den ProPublica-Daten (grüner Punkt) bei eingezeichnetem maximalen PPV_k (obere Kurve, blau) und minimalem PPV_k (untere Kurve, rot) für jeweils fixierte AUC auf der x-Achse, Abbildung aus (Krafft 2017).

Bewertungsmaßstab in der ADM-gestützten Kriminalprognose blieb deutlich hinter den Erwartungen zurück. Die Abweichung vom näher am richterlichen Entscheidungsprozess evaluierenden PPV_k kann inakzeptabel hoch sein, sodass eine zukünftige Heranziehung der *AUC* als ausschließliches Qualitätsmaß kritisch zu hinterfragen ist.

Sollte die *AUC* weiterhin in der Justiz Verwendung finden, müssten Gutachter wie Richter die Aussagekraft eines hohen *AUC*-Wertes richtig zu interpretieren wissen und dürften die Fähigkeit der Instrumente zur Rückfälligkeitsprognose aufgrund des Vorliegens hoher Validitätswerte nicht überschätzen (Eher et al. 2008). Hier bedarf es massiver Aufklärung, da die Tendenz besteht, dass Menschen softwarebasierte Prognosen als verlässlicher und objektiver empfinden, was die Gefahr einer unkritischen Übernahme dieser Prognosen birgt.

Die aufgeführten Probleme haben weiterhin gezeigt, dass es von immenser Bedeutung ist, das Anwendungsgebiet sowie die dort vorherrschende Datenlage und Qualität genauestens mit der Lernumgebung der Algorithmen abzugleichen, vgl. Burnham et al. (2002), denn „der beste Klassifikationsalgorithmus ist gerade so gut wie die ihm vorliegende Information" (Hengen et al. 2004).

Mögliche Konsequenzen und Forderungen

Sollte im deutschen Justizwesen die Einführung ADM-gesteuerter Prozesse zur Diskussion stehen, könnte Deutschland von den Erfahrungen und Fehlern anderer Länder profitieren. Es ist zu hoffen, dass dies ohne überstürzten politischen Aktionismus, sondern mit Bedacht nach einer ausführlichen Debatte erfolgt. Keinesfalls sollte es wie beim Jugendstrafrecht in den USA (Baird 2009, S. 2) zu einer voreiligen Nutzung von Algorithmen kommen. Bei der Risikoprognose steht ein Mensch im Mittelpunkt algorithmischer Betrachtung, sodass Auswahl, Überprüfung und Nutzung von algorithmischen Entscheidungshilfen größte wissenschaftliche und gesellschaftliche Aufmerksamkeit geschenkt werden muss. Es geht um existenzielle Urteile für die Betroffenen und Fehlurteile könnten fatale Auswirkungen auf deren Leben haben[5].

Damit aber eine weiterführende Forschung betrieben werden kann, müsste der Staat die notwendigen finanziellen Mittel zur Verfügung stellen. Es ist kein akzeptabler Status quo, wenn kritische Untersuchungen zu Risiken und gesellschaftlichen Folgen der ADM-Prozesse abhängig vom Interesse und verfügbaren Budget beliebiger Institutionen sind. So ist der Anstoß der Fairness-Debatte um das COMPAS-Tool nur dem Engagement des Recherchebüros ProPublica zu verdanken.

Schon normative Entscheidungen, z. B. über Fairness-Kriterien, müssten bei der Gestaltung eines ADM-Prozesses im Konsens mit der Gesellschaft gefällt werden. Sinnvolle Handlungsempfehlungen hierfür gibt das ADM-Manifest (AlgorithmWatch 2016) in den Punkten 3/4:

3. ADM-Prozesse müssen nachvollziehbar sein, damit sie demokratischer Kontrolle unterworfen werden können.
4. Demokratische Gesellschaften haben die Pflicht, diese Nachvollziehbarkeit herzustellen: durch eine Kombination aus Technologien, Regulierung und geeigneten Aufsichtsinstitutionen.

Demzufolge wären die Vertreiber gefordert, die Algorithmen ihrer angebotenen Tools so weit offenzulegen, „dass Erklärbarkeit, Nachvollziehbarkeit, unabhängige Überprüfbarkeit und die Möglichkeiten zur forensischen Datenanalyse gegeben sind" (Lischka und Klingel 2017). Sollte dies nicht geschehen, wäre es auch hier Aufgabe des Staates, durch geeignete Maßnahmen eine informierte Debatte zu ermöglichen.

Abschließend lässt sich konstatieren, dass die Gesellschaft bei der Integration eines Risikovorhersage-Instruments die Auswahl des Bewertungsmaßstabs als eine der wichtigsten Modellierungsentscheidungen verstehen muss, denn:

> *„We recognize that creation of valid, reliable, and robust risk assessment instruments is both a science and an art."*[6]

Referenzen

Aggarwal CC (2015) Data mining; The textbook. Springer, Cham

AlgorithmWatch (2016) Das ADM-Manifest | The ADM Manifesto. https://algorithmwatch.org/das-adm-manifest-the-adm-manifesto/. Zugegriffen: 22. Februar 2018

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias; There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, 23. Mai 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Zugegriffen: 22. Februar 2018

Azariadis C (1981) Self-fulfilling prophecies. J Econ Theory 25:380–396. doi:10.1016/0022-0531(81)90038-7

Baird C (2009) A question of evidence; A critique of risk assessment models used in the justice system. National Council on Crime and Delinquency, Madison, WI

Barnoski R, Drake EK (2007) Washington's Offender Accountability Act; Department of Corrections' static risk instrument. Washington State Institute for Public Policy, Olympia, WA. http://www.wsipp.wa.gov/ReportFile/977/Wsipp_Washingtons-Offender-Accountability-Act-Department-of-Corrections-Static-Risk-Instrument_Full-Report-Updated-October-2008.pdf

Beizer B (1995) Black-box testing; Techniques for functional testing of software and systems. Wiley, New York, NY

Bliesener T, Lösel F, Köhnken G (Hrsg) (2014) Lehrbuch der Rechtspsychologie. Huber, Bern

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 30:1145–1159. doi:10.1016/S0031-3203(96)00142-2

Brennan T, Dieterich B, Breitenbach M, Mattson B (2009a) A Response to "Assessment of Evidence on the Quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)". Northpointe Institute for Public Management, Inc. http://www.northpointeinc.com/files/whitepapers/Response_to_Skeem_Louden_Final_071509.pdf

Brennan T, Dieterich W, Ehret B (2009b) Evaluating the predictive validity of the Compas risk and needs assessment system. "Crim Justice Behav 36:21–40. doi:10.1177/0093854808326545

Burnham BR, Thompson DF, Jackson WG (2002) Positive predictive value of a health history questionnaire. Mil Med 167:639–642. doi:10.1093/milmed/167.8.639

Cesare B (1764) Dei delitti e delle pene. In: Opera immortale del Marchese di Beccaria. R. Sammer, Wien, 1798

Chettiar IM, Gupta V (2011) Smart reform is possible; States reducing incarceration rates and costs while protecting communities. American Civil Liberty Union (ACLU). SSRN, 27. September 2011. doi:10.2139/ssrn.1934415

Danziger S, Levav J, Avnaim-Pesso L (2011) Extraneous factors in judicial decisions. PNAS USA 108:6889–6892. doi:10.1073/pnas.1018033108

Diakopoulos N (2014) Algorithmic accountability reporting; On the investigation of black boxes. Tow Center for Digital Journalism, Columbia University, Februar 2014. http://towcenter.org/wp-content/uploads/2014/02/78524_Tow-Center-Report-WEB-1.pdf

Dieterich W, Mendoza C, Brennan T (2016) COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Northpointe. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Dittmann V (2003) Was kann die Kriminalprognose heute leisten? In: Häßler F, Rebernik E, Schnoor K, Schläfke D, Fegert JM (Hrsg) (2003) Forensische Kinder-, Jugend- und Erwachsenenpsychiatrie; Aspekte der forensischen Begutachtung. Schattauer, Stuttgart, S 173–187

Döbele A-L (2014) Standardisierte Prognoseinstrumente zur Vorhersage des Rückfallrisikos von Straftätern; Eine kritische Betrachtung des Einsatzes in der Strafrechtspflege aus juristischer Sicht. Kovač, Hamburg

Eher R, Rettenberger M, Schilling F, Pfäfflin F (2008) Validität oder praktischer Nutzen? Rückfallvorhersagen mittels Static-99 und SORAG. Eine prospektive Rückfallstudie an 275 Sexualstraftätern. Recht & Psychiatrie 26:79–88

Electronic Privacy Information Centre, EPIC (2017) Algorithms in the criminal justice system. https://epic.org/algorithmic-transparency/crim-justice/. Zugegriffen: 25. Januar 2018

Endrass J, Urbaniok F, Held L, Vetter S, Rossegger A (2008) Accuracy of the Static-99 in predicting recidivism in Switzerland. Int J Offender Ther Comp Criminol. doi:10.1177/0306624X07312952

Farabee D, Zhang S, Roberts REL, Yang J (2010) COMPAS validation study; Final report. UCLA Integrated Substance Abuse Programs (ISAP), 15. August 2010. https://www.cdcr.ca.gov/adult_research_branch/Research_Documents/COMPAS_Final_report_08-11-10.pdf

Fergusson DM, Fifield JK, Slater SW (1977) Signal detectability theory and the evaluation of prediction tables. J Res Crime Delinq 14:237–246. doi:10.1177/002242787701400209

Guy LS (2008) Performance indicators of the structured professional judgment approach for assessing risk for violence to others; A meta-analytic survey. Dissertation, Simon Fraser University

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36. doi:10.1148/radiology.143.1.7063747

Hengen H, Feid M, Pandit M (2004) Überwacht lernende Klassifikationsverfahren im Überblick, Teil 1 (Overview of Supervised learning Classification Methods, Part 1). Automatisierungstechnik/Methoden und Anwendungen der Steuerungs-, Regelungs- und Informationstechnik 52(3):A1–A8. doi:10.1524/auto.52.3.A1.34763

Krafft TD (2017) Qualitätsmaße binärer Klassifikatoren im Bereich kriminalprognostischer Instrumente der vierten Generation. Masterarbeit, Fachbereich Informatik, TU Kaiserslautern. arXiv:1804.01557

Lansing S (2012) New York State COMPAS-probation risk and need assessment study; Examining the recidivism scale's effectiveness and predictive accuracy. NCJ 247345. https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=269445.

Leushuis E, van der Steeg JW, Steures P, Bossuyt PMM, Eijkemans MJC, van der Veen F, Mol BWJ, Hompes PGA (2009) Prediction models in reproductive medicine; A critical appraisal. Hum Reprod Update 15:537–552. doi:10.1093/humupd/dmp013

Lischka K, Klingel A (2017) Wenn Maschinen Menschen bewerten. Bertelsmann Stiftung. doi:10.11586/2017025

Maschke W (2008) Die Kriminalprognose im Einzelfall. In: Dölling D (Hrsg) (2009) Gutachten im Jugendstrafverfahren. DVJJ, Heidelberg, S 85–102

Nedopil N, Groß G (2005) Prognosen in der Forensischen Psychiatrie; Ein Handbuch für die Praxis. Pabst, Lengerich, Westf.

Northpointe (2012a) COMPAS risk & need assessment system; Selected questions posed by inquiring agencies. http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf. Zugegriffen: 22. Februar 2018

Northpointe (2012b) Practitioner's guide to COMPAS core. http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf

Obergfell-Fuchs J (2011) Gefährliche Straftäter aus kriminologischer und psychologischer Sicht. In: Sicherungsverwahrung und Führungsaufsicht; Wie gehen wir mit gefährlichen Straftätern um? Evangelische Akademie, Bad Boll

Peterson W, Birdsall T, Fox W (1954) The theory of signal detectability. Trans IRE Prof Group Inf Theory 4(4):171–212. doi:10.1109/TIT.1954.1057460

Rettenberger M, von Franqué F (Hrsg) (2013) Handbuch kriminalprognostischer Verfahren. Hogrefe, Göttingen, Bern, Wien

Statista (2017) Länder mit der größten Anzahl an Inhaftierten (Februar 2017*). https://de.statista.com/statistik/daten/studie/3212/umfrage/laender-mit-den-meisten-gefangenen-im-jahr-2007/. Zugegriffen: 22. Februar 2018

US Supreme Court (1966) Baxstrom v. Herold, 383 U.S. 107 (1966). https://supreme.justia.com/cases/federal/us/383/107/case.html

Wisconsin Department of Correction (2018) COMPAS. https://doc.wi.gov/Pages/AboutDOC/COMPAS.aspx. Zugegriffen: 29. März 2018

Zweig KA (2016) 2. Arbeitspapier: Überprüfbarkeit von Algorithmen. AlgorithmWatch, 7. Juli 2016. http://algorithmwatch.org/zweites-arbeitspapier-ueberpruefbarkeit-algorithmen/. Zugegriffen: 22. Februar 2018

Zweig KA, Wenzelburger G, Krafft TD (2018) On chances and risks of security related algorithmic decision making systems. Erscheint in European Journal for Security Research

Anmerkungen

1 *Die Ausführungen zur Handhabung der Risikoprognostik im deutschen Justizwesen sind in Kapitel 1.2.2 meiner Masterarbeit (Krafft 2017) nachzulesen.*

2 *Eine genauere Erklärung der Terminologie ist im Kapitel 2 meiner Masterarbeit (Krafft 2017) nachzulesen.*

3 *Die mathematische Beweisführung kann bei Interesse in Kapitel 4 der Masterarbeit (Krafft 2017) nachgelesen werden.*

4 *Der Wert für die tatsächliche Rückfälligkeit eines Individuums bezieht sich in dem durch ProPublica offerierten Datensatz auf ein Zweijahresfenster, soweit es das Broward County Sheriff's Office in Florida erfassen konnte.*

5 *Als Beispiel sei der ‚Feedback Loop' genannt, nach dem ein ‚false positive' (fälschlicherweise Verurteilter) tatsächlich im Sinne der ‚self fulfilling prophecy' (Azariadis 1981) kriminell würde.*

6 *Wir stellen fest, dass die Erstellung eines fundierten, zuverlässigen und robusten Risikobeurteilungsinstruments sowohl eine Wissenschaft als auch eine Kunst ist. (Baird 2009, S. 10)*