Faires Machine Learning

Eine Open Educational Resource

Autor:innen: Fiona Fischer, Linda Fernsel
im Projekt “Fair Enough?” (16DHB4002)
der Forschungsgruppe Informatik und Gesellschaft,
an der Hochschule für Technik und Wirtschaft (HTW) Berlin

Version: November 2023

Inhaltsverzeichnis

1. Über diese OER
2. Einleitung
3. Anwendungsfälle von Gesichtserkennungstechnologie
4. Machine Learning Modelle kritisch betrachten
5. Zusammenfassung & Fazit
6. Quellen

1. Über diese OER

Inhalte, Lernziele, Voraussetzungen und Begleitmaterial dieser OER.

Diese OER bietet Struktur, Inhalte und Aufgabenideen für eine Lehreinheit zum Thema *Faires Machine Learning*. Die Einheit behandelt sowohl soziale als auch technische Aspekte von fairem Machine Learning und soll so Personen, die in Zukunft möglicherweise mit Machine-Learning-Modellen in Berührung kommen, befähigen, sich kritisch mit diesen auseinanderzusetzen.

Die Anwendung von Machine Learning in Gesichtserkennungstechnologien steht bei dieser OER im Vordergrund - alle Methoden werden an Hand des erfundenen Gesichtserkennungstools *IdentiTOP* demonstriert.

Diese OER verfolgt folgende Lernziele:

- Die Teilnehmenden kennen Anwendungsfälle für Gesichtserkennungstechnologie.
- Die Teilnehmenden können gängige Risiken bei der Anwendung von Machine-Learning-Modellen identifizieren (Soziale Perspektive).
- Die Teilnehmenden kennen Qualitätsmetriken für Machine Learning Modelle (Technologische Persoektive).
- Die Teilnehmenden können zu Risiken passende Qualitätsmetriken auswählen (Technologische Perspektive.

- Die Teilnehmenden können Qualitätsmetriken basierend auf der Wahrheitsmatrix berechnen (Technologische Perspektive).
- Die Teilnehmenden können die Bedeutung der Qualitätsmessergebnisse für die Risiken von Machine Learning Modellen interpretieren (Technologische Perspektive + Soziale Perspektive).

Diese OER kann das Thema “Machine Learning”, was z.B. im Oberstufeninformatikunterricht oder in der Hochschule unterrichtet wird, ergänzen. Es wird daher vorausgesetzt, dass die Teilnehmenden bereits mit einigen Grundlagen von “Machine Learning” vertraut sind. Die Verwendung der ergänzenden Jupyter Notebooks setzt außerdem ein grundlegendes Verständnis für Programmierung voraus.

Python mit Jupyter Notebook kann verwendet werden um auch die programmatische Durchführung von Audits zu erlernen. Die Jupyter Notebooks sind im Repository unter https://gitlab.com/iug-research/oer-faires-machine-learning (bzw. verkürzt https://bit.ly/identitop) verfügbar.

- An Hand des Notebooks fairness-audit-identitop.ipynb kann der Audit von *IdentiTOP* programmatisch nachvollzogen werden.
- Das Notebook fairness-audit-kreditwuerdigkeit-aufgabe.ipynb enthält eine weiterführende Aufgabe, bei der ein eigener Audit angeleitet durchgeführt werden soll.
- Die Lösung ist im Notebook fairness-audit-kreditwuerdigkeit-loesung.ipynb enthalten.

Google Colab bietet die Möglichkeit, Jupyter Notebooks mit nur einem Browser nutzen zu können.

Lernende können diese Lerneinheit auch eigenständig lernen. Dafür gibt es ein Handout unter https://iug.htw-berlin.de/wp-content/uploads/2023/11/Handout-Faires-Machine-Learning.pdf, welches in Kombination mit dem Jupyter Notebook zu IdentiTOP zu verwenden ist.

2. Einleitung

Dieser Abschnitt führt Thema und Begriffe ein und erzeugt Motivation, sich mit dem Thema zu beschäftigen.

Mit Hilfe von Machine Learning können wir Modelle entwickeln, die Daten anhand von Eigenschaften klassifizieren oder darüber Vorhersagen treffen. Es hat sich jedoch herausgestellt, dass diese Modelle nicht immer fair sind.

Das zeigt zum Beispiel das Projekt “Gender Shades” [4]. Bei diesem Projekt wurden automatisierte Gesichtsanalysedienste hinsichtlich ihrer Fairness bezüglich Geschlecht und Hautfarbe untersucht [4]. Die Analyse der Gesichtsanalysealgorithmen ergab, dass Frauen im Vergleich zu Männern und dunkelhäutige Personen im Vergleich zu hellhäutigen Personen öfter falsch klassifiziert wurden [4]. Dunkelhäutige Frauen wurden dabei am

häufigsten falsch klassifiziert [4]. Die Wissenschaftler:innen stellten fest, dass dies vor allem an den Trainingsdaten der Algorithmen lag: Bestehende Datensätze umfassten überwiegend hellhäutige Männer [4].

Daher hat das Themengebiet “Faires Machine Learning” zum Ziel, Machine Learning aus verschiedenen Perspektiven kritisch zu betrachten. Dabei wird ergründet, wie fair Machine-Learning-Systeme sind (Stufe 1), was Ursachen für Unfairness sind (Stufe 2) und wie Unfairness behoben werden kann (Stufe 3).

Die Fairness eines Modells lässt sich aus verschiedenen Blickpunkten betrachten:

1. Die sozialen Perspektive betrifft die Abschätzung sozialer Folgen und Risiken eines Modells wenn Menschen mit dem Modell interagieren und es verwenden.
2. Die technischen Perspektive betrifft die technische Messung der Modellqualität nach mathematischen Definitionen.

3. Anwendungsfälle von Gesichtserkennungstechnologie

Dieser Abschnitt gibt einen Überblick über mögliche Anwendungsfälle von Gesichtserkennungstechnologie. Er stellt außerdem den Anwendungsfall “IdentiTOP” vor, der im weiteren Verlauf der OER als Beispiel dient.

Gesichtserkennungtechnologie verwendet Machine Learning Modelle für drei verschiedene Level von Aufgaben im Bezug auf das Gesicht einer Person [3]:

1. Gesichtserkennung: Erkennen, ob sich ein Gesicht im Bild befindet.
2. Eigenschaftenerkennung: Erkennen um was für ein Gesicht es sich handelt.
3. Identitätserkennung: Erkennen, wessen Gesicht es ist.

Beispiele für Anwendungsfälle:

Gesichtserkennung	**Eigenschaftenerkennung**	**Identitätserkennung**
Foto-Filter bei Apps wie SnapChat [1]	Geschlecht und Alter erkennen um zur Zielgruppe passende Werbung vorzuschlagen [2]	Identifikation polizeilich gesuchter Personen bei Videoüberwachung [5]
Überwachung ob bei einer Online-Prüfung eine Person alleine vor dem PC ist [8]	Interesse im Unterricht messen als Feedback für Lehrkräfte [7]	Identifikation von Teilnehmenden bei einer Online-Prüfung [6]

Anwendungsfall *IdentiTOP* - ein fiktives System zur Identifikation von Teilnehmenden bei einer Online-Prüfung

Online-Prüfungen bergen die Gefahr, dass Lernende die Klausur nicht selbst schreiben, sondern jemand anderes damit beauftragen. Das Ziel von *IdentiTOP* ist es, diese Art von Betrug bei Online-Prüfungen zu verhindern.

Wenn eine Person sich bei einer Lehreinrichtung (z.B. Uni) registriert, muss sie eine Ausweiskopie einreichen. Die Ausweisbilder der Personen werden in einer Datenbank gespeichert. *IdentiTOP* hat Zugriff auf die Datenbank der Ausweisbilder aller Lernenden und weiß außerdem, wer welchen Kurs besucht. Bei der Registrierung an der Lehreinrichtung und zu Beginn der Online-Prüfung werden Prüflinge über den Einsatz von *IdentiTOP* und die Konsequenzen bei Betrug informiert.

Während der Prüfung nimmt *IdentiTOP* an zufälligen Zeitpunkten mit der Webcam ein Bild von der Person auf, die die Prüfung schreibt. Das System holt sich dann aus der Datenbank das Ausweisbild der angemeldeten Person. Dann gibt es das Ausweisbild zusammen mit dem Webcam-Foto in ein Machine Learning Modell ein, was vorhersagt, ob es sich bei den Personen auf den Fotos um dieselbe Person handelt.

Falls es sich um dieselbe Person handelt, kann die Person ungestört weiterarbeiten und wird in einigen Minuten erneut von *IdentiTOP* überprüft.

Falls es sich jedoch laut dem Modell nicht um dieselbe Person handelt, wird der Prüfungsversuch vom System abgebrochen und die betrügende Person erhält für den Prüfungsversuch ein “ungenügend”.
Die Hersteller:innen von *IdentiTOP* werben damit, dass Aufsichtskräfte während der Prüfung nicht mehr alle (möglicherweise hunderte) Prüflinge per WebCam ständig im Blick behalten müssten und dass das System den Aufsichtskräften das unangenehme persönliche Eingreifen bei Betrug ersparte. So könnte Betrug konsequent erkannt und geahndet werden. Betrug würde außerdem reduziert, da Prüflinge über Einsatz und Konsequenzen informiert werden.

Unterrichtsidee: Fragen Sie die Teilnehmenden, ob sie schonmal mit Gesichtserkennungstechnologie in Berührung gekommen sind, und wozu Gesichtserkennung in diesem Fall eingesetzt wurde. Sammeln Sie die Anwendungsfälle auf Papier an der Tafel. Sortieren Sie die Anwendungsfälle in die drei Stufen der Gesichtserkennungstechnologie ein.

4. Machine Learning Modelle kritisch betrachten

Dieser Abschnitt erklärt die Schritte, wie Machine Learning Modelle kritisch betrachtet werden können.

Die Kritische Betrachtung von Machine Learning Modellen auf Stufe 1 erfolgt in vier Schritten: (1) Identifizierung der Modellrisiken, (2) Auswahl der Qualitätsmetriken, (3)

Berechnung der Qualitätsmetriken und (4) Interpretation der Qualitätsmetriken. Zum Schluss kann man noch einen weiteren Schritt anschließen, der jedoch bereits Stufe 2 und 3 der Kritischen Betrachtung anschneidet: (5) Erarbeiten von Verbesserungsvorschlägen.

4.1. Identifizierung der Risiken eines Modells

Mit Hilfe der sozialen Perspektive können die Anwendungsfall-bezogene sozialer Folgen und Risiken eines Machine Learning Modells, wenn Menschen das Modell verwenden, identifiziert werden.
Folgende Fragen können helfen, Risiken zu identifizieren:

- Was sind mögliche Nachteile und negative Folgen eines Systems?
- Welche Fehler können auftreten?
- Wie wirken sich diese Probleme auf Personen aus?
- Können diese Probleme strukturell auftreten und Diskriminierung verursachen?

Beispiele für Diskussionen von Anwendungsfällen für Gesichtserkennungstechnologie:

Foto-Filter bei Apps wie SnapChat [1] Die Verwendung von Gesichtserkennungstechnologie in Apps wie SnapChat ermöglicht es Benutzer:innen, interaktive, unterhaltsame oder verschönernde Filter auf ihre Gesichter anzuwenden. Es können jedoch Fehler bei der Erkennung von Gesichtsmerkmalen auftreten, was zu unerwünschten Filtereffekten führt. Außerdem könnte der Gebrauch verschönernder Filter das eigene Selbstbild und das der anderen Nutzer:innen stören.
Geschlecht und Alter erkennen um zur Zielgruppe passende Werbung vorzuschlagen [2] Das Erkennen von Geschlecht und Alter ermöglicht eine gezielte Werbung, die potentiell relevanter für die angesprochene Person ist. Falsche Erkennung von Geschlecht oder Alter könnte zu irrelevanten Werbeanzeigen führen. Dies könnte außerdem psychische Folgen haben, z.B. bei Trans-Personen. Auch bei korrekter Erkennung könnten Personen sich missverstanden oder stereotypisiert fühlen. Es ist wichtig, sicherzustellen, dass keine Stereotypen verstärkt werden und dass Personen die Möglichkeit haben, ihre Präferenzen manuell anzugeben.
Identifikation polizeilich gesuchter Personen bei Videoüberwachung [5] Die Identifikation polizeilich gesuchter Personen könnte die öffentliche Sicherheit erhöhen und die Polizeiarbeit effizienter machen. Falsche Identifikationen könnten unschuldige Bürger:innen ins Visier der Strafverfolgung bringen. Unschuldige könnten festgenommen oder belästigt werden, was zu psychischen und körperlichen Problemen führen kann [10], sowie zu einem Vertrauensverlust in die Strafverfolgungsbehörden. Die Technologie sollte also mit großer Vorsicht und in Kombination mit menschlicher Überprüfung eingesetzt werden.
Überwachung der alleinigen Anwesenheit einer Person bei einer Online-Prüfung am PC [8] Die Überwachung, ob bei einer Online-Prüfung eine Person alleine vor dem PC ist, soll sicherstellen, dass Betrug bei Prüfungen verhindert wird. Die Technologie könnte jedoch

andere Objekte fälschlicherweise als zweite Person interpretieren, oder wiederum andere Personen nicht als solche erkennen. Prüflinge könnten ungerechterweise des Betrugs beschuldigt werden, was schwerwiegende akademische Konsequenzen haben könnte. Außerdem könnte das Gefühl, überwacht zu werden, Prüflinge unter Druck setzen, wodurch sie sich anders verhalten als gewöhnlich.
Interesse im Unterricht messen als Feedback für Lehrkräfte [7] Wenn das Interessiertheit der Lernenden im Unterricht gemessen wird, könnten Lehrpersonen wertvolles Feedback erhalten, um ihren Unterricht anzupassen und zu verbessern. Nicht alle Gesichtsausdrücke spiegeln jedoch das tatsächliche Interesse oder das Verständnis wider. Dies betrifft z.B. Personen aus unterschiedlichen Kulturkreisen, die unterschiedliche Mimik haben, oder Personen, die ihre Gefühle auf Grund von anderen inhärenten Eigenschaften anders zum Ausdruck bringen als andere. Lernende könnten ungerechterweise als desinteressiert oder abgelenkt eingestuft werden, was zu ungerechtfertigten pädagogischen Maßnahmen führen könnte. Diese Technologie sollte also nicht als alleiniges Mittel zur Bewertung des Interesses verwendet werden.
Identifikation von Teilnehmenden bei einer Online-Prüfung [6] Die Identifikation von Teilnehmenden bei einer Online-Prüfung stellt sicher, dass die richtige Person die Prüfung ablegt. Eine falsche Erkennung könnte dazu führen, dass eine Person ungerechterweise von einer Prüfung ausgeschlossen würde, was zu akademischen und emotionalen Belastungen führen könnte.

Unterrichtsidee: Lassen Sie die Lernenden die zuvor zusammen getragenen Anwendungsfälle diskutieren, um das Identifizieren von Risiken zu üben.

Unterrichtsidee: Das Thema “Gesichtserkennungstechnologie” lässt sich gut mit Exkursen zur Geschichte und zu den sozialen Folgen von Überwachung, sowie mit weiterführenden Diskussionen zu Datenschutz verbinden.

Risiken von *IdentiTOP*:

- (R1) Wenn *IdentiTOP* fälschlicherweise annimmt, dass es sich nicht um dieselbe Person handelt, also die Person nicht identifiziert werden kann, könnte das zu hoher emotionaler Belastung bei der unschuldigen Person führen, sowie organisatorischen Aufwand (um dem Urteil durch IdentiTOP zu widersprechen und einen neuen Prüfungsversuch zu erhalten) oder gar akademischen Nachteilen (falls der Person kein neuer Versuch gewährt wird).
- (R2) Wenn *IdentiTOP* fälschlicherweise annimmt, dass es sich um dieselbe Person handelt, obwohl die Personen sich lediglich ähnlich sehen, würde der Betrug unerkannt bleiben und die wahre Leistungsfähigkeit der Person könnte nicht eingeschätzt werden. Die Person könnte also in einen Aufbaukurs aufgenommen werden, oder einen Job erhalten, wofür sie gar nicht qualifiziert ist, was sich je nach genauer Situation negativ auf alle Beteiligten auswirken würde.
- (R3) Falls *IdentiTOP* für hellhäutige Männer besser funktioniert, als für andere Personengruppen (siehe “Gender Shades” [4]), werden manche Gruppen stärker

benachteiligt als andere und bestehende strukturelle Diskriminierung wird verstärkt.
- (R4) Falls IdentiTOP für Personen mit besserer Webcam besser funktioniert, als für andere, werden finanziell schlechter gestellte Personengruppen diskriminiert.
- (R5) Prüflinge könnten sich in ihrer Privatsphäre gestört fühlen, da IdentiTOP unvorhergesehen Fotos von ihnen macht. Das könnte zu emotionaler Belastung und damit einem schlechteren Prüfungsergebnis führen.

4.2. Qualitätsmetriken auswählen

Wir haben in der Betrachtung der sozialen Perspektive festgestellt, dass jedes Modell individuelle Risiken mit sich bringt. Es ist daher notwendig, solche Metriken auszuwählen, die helfen, die spezifischen Risiken einzuschätzen. Dafür muss zu jedem Risiko die passende Metrik gefunden werden. Viele Metriken basieren auf der Wahrheitsmatrix (en: confusion matrix).

Die Wahrheitsmatrix

Eine Wahrheitsmatrix gibt einen Überblick über die Richtigkeit der Ergebnisse eines Modells, indem sie die Modellvorhersagen mit der Wirklichkeit (also dem richtigen Ergebnis) vergleicht. Die Ergebnisse eines Machine Learning Modells können dabei in “positiv” und “negativ” eingeteilt werden.

Die Wahrheitsmatrix enthält die Anzahl richtiger Vorhersagen: die Anzahl richtig positiver Fälle (en: true positive; kurz: **TP**) und die Anzahl richtig negative Fälle (true negative; kurz: **TN**). Außerdem enthält sie die Anzahl falscher Vorhersagen: die Anzahl falsch positiver Fälle (en: false positive; kurz: **FP**) und die Anzahl falsch negativer Fälle (true negative; kurz: **FN**).

Konkret ist die Wahrheitsmatrix nach folgendem Schema aufgebaut [11]:

Vorhersage → Wirklichkeit ↓	Als positiv vorhergesagt	Als negativ vorhergesagt
Wirklich positiv	TP	FN
Wirklich negativ	FP	TN

Eine hohe Modellqualität ist gegeben, wenn TP und TN hoch und FP und FN niedrig sind.

Wahrheitsmatrix für *IdentiTOP*
Das Ergebnis “beide Bilder zeigen dieselbe Person” zählt als positives Ergebnis und das Ergebnis “beide Bilder zeigen unterschiedliche Personen” zählt als negatives Ergebnis.

IdentiTOP wurde gezielt getestet. 160 der 300 Testpersonen haben “betrogen”, damit geprüft werden kann, wie gut *IdentiTOP* funktioniert.

Dies ist die Wahrheitsmatrix für *IdentiTOP*:

	Als positiv vorhergesagt	Als negativ vorhergesagt
Wirklich positiv	113	27
Wirklich negativ	10	150

Dann sagt diese Wahrheitsmatrix aus, dass:

- 113 mal richtig vorhergesagt wurde, dass es sich auf beiden Bildern um dieselbe Person handelt
- 10 mal falsch vorhergesagt wurde, dass es sich auf beiden Bildern um dieselbe Person handelt
- 27 mal falsch vorhergesagt wurde, dass es sich auf beiden Bildern nicht um dieselbe Person handelt
- 150 mal richtig vorhergesagt wurde, dass es sich auf beiden Bildern nicht um dieselbe Person handelt

Das Modell hat also in 37 von 300 Fällen einen Fehler gemacht und lag in den restlichen 263 Fällen richtig.

Unterrichtsidee: Stellen Sie den Lernenden die obigen Aussagen der Wahrheitsmatrix als Fragen, z.B.: “Wie oft wurde richtig vorhergesagt, dass es sich auf beiden Bildern um dieselbe Person handelt?”

Welche Qualitätsmetriken gibt es?

Gängige Metriken zur Prüfung der Modellqualität sind die Richtigkeit (en: accuracy), Genauigkeit (en: precision) oder Trefferquote (en: recall) [9]. Die Richtigkeit bezieht sich auf alle Vorhersagen (TP + TN + FP + FN), die Genauigkeit nur auf die Fälle mit positiver Vorhersage (TP + FP) und die Trefferquote bezieht sich auf die wirklich positiven Fälle (TP + FN). Diese Metriken werden wie folgt anhand der Wahrheitsmatrix berechnet [11]:

- Richtigkeit = (TP + TN) / (TP + TN + FP + FN)
- Genauigkeit = TP / (TP + FP)
- Trefferquote = TP / (TP + FN)

Berechnung gängiger Metriken für *IdentiTOP*

- Richtigkeit = 263 / 300 ≈ 0,88
- Genauigkeit = 113 / 123 ≈ 0,92
- Trefferquote = 113 / 140 ≈ 0,81

Unterrichtsidee: Fordern Sie die Lernenden auf, die Metriken basierend auf der Wahrheitsmatrix und den Metrikdefinitionen zu berechnen.

Zusätzlich gibt es weniger gängige Metriken, z.B. die **Spezifität** (en: specificity), die Falsch-negativ-Rate (en: false negative rate; kurz: **FNR**), die Falsch-positiv-Rate (en: false

positive rate; kurz: **FPR**), die Falscherkennungsrate (en: false discovery rate; kurz: **FDR**) und die Falschauslassungsrate (en: false omission rate; kurz: **FOR**), die auf der Wahrheitsmatrix basieren [11]. Diese Metriken werden wie folgt an Hand der Wahrheitsmatrix berechnet [11]:

- Spezifität = TN / (TN + FP)
- FNR = FN / (FN + TP)
- FPR = FP / (FP + TN)
- FDR = FP / (FP + TN)
- FOR = FN / (FN + TN)

Fairness

Fairness kann man mit Hilfe mathematischer Fairness-Definitionen messen.
Fairness-Definitionen sagen aus, dass ein Modell für verschiedene **Gruppen gleich gut** funktioniert.
“Gruppen” meint die Unterteilung von Personen nach verschiedenen Attributen, z.B. “Geschlecht”. Eine Gruppe wäre in diesem Fall “Frauen”. Gruppen können auch “intersektional” sein. Das bedeutet, dass Personen nach mehreren Attributen aufgeteilt wurden, z.B. “Geschlecht” und “Hautfarbe”. In diesem Fall wäre z.B. “dunkelhäutige Frauen” eine Gruppe.
“Gleich” meint, dass die Qualität des Modells soll für alle Gruppen ähnlich sein. Maße für Ähnlichkeit zweier Zahlen sind die Differenz (z.B. FPR von Gruppe 1 minus FPR von Gruppe 2) oder das Verhältnis (z.B. FPR von Gruppe 1 geteilt durch FPR von Gruppe 2).
“Gut” meint die messbare Qualität des Modells, z.B. die FPR. Die Qualitätsmetrik wird je nach Risiko ausgewählt.

Einige benannte Fairness-Definitionen sind [13]:

- Vorhersageparität (en: Predictive Parity), das entspricht der Genauigkeitsgleichheit. Das heißt, die Genauigkeit für jede Gruppe muss gleich hoch sein.
- Chancengleichheit (en: Equal Opportunity), das entspricht der Falsch-negativ-Raten-Gleichheit (FNR-Gleichheit). Das heißt, die FNR für jede Gruppe muss gleich niedrig sein.
- Prediktive Gleichheit (en: Predictive Equality), das entspricht der Falsch-positiv-Raten-Gleichheit (FPR-Gleichheit). Das heißt, die FPR für jede Gruppe muss gleich niedrig sein.

Auswahl der Qualitätsmetriken für *IdentiTOP*:

- (R1 - *IdentiTOP* nimmt fälschlicherweise an, dass es sich nicht um dieselbe Person handelt) ist akzeptabel, falls die wirklich positiven Fälle wenige negative Vorhersagen erhalten (FN). Dies kann mit FNR berechnet werden. Bei einer niedrigen FNR ist es weniger wahrscheinlich, dass R1 eintritt.
- (R2 - *IdentiTOP* nimmt fälschlicherweise an, dass es sich um dieselbe Person handelt) ist akzeptabel, falls alle tatsächlich negativen Fälle wenige falsch-positive Vorhersagen (FP) erhalten. Das kann mit FPR berechnet werden. Bei einer niedrigen FPR ist es weniger wahrscheinlich, dass R2 eintritt.
- (R3 - *IdentiTOP* funktioniert für hellhäutige Männer besser, als für andere Personengruppen) besteht aus zwei Teilen:

    - Risiko a) ist akzeptabel, wenn das Qualitätskriterium für (R1) für Personen aller Geschlechter und Hautfarben gleichermaßen zutrifft. Die zugehörige Fairness-Definition ist die FNR-Gleichheit, bzw. Chancengleichheit.
    - Risiko b) ist akzeptabel, wenn das Qualitätskriterium für (R2) für Personen aller Geschlechter und Hautfarben gleichermaßen zutrifft. Die zugehörige Fairness-Definition ist die FPR-Gleichheit, bzw. Prädiktive Gleichheit.
- (R4) entspricht (R3), allerdings unterteilen sich die Gruppen hier nach Webcamqualität. Die Behandlung dieses Risikos lassen wir in diesem Beispiel aus, da es (R3) ähnelt und nicht gesondert demonstriert werden muss. In Wirklichkeit wäre es aber natürlich sehr wichtig, R4 auch zu prüfen.
- (R5) ist akzeptabel, wenn die meisten Prüflinge aussagen, sich durch *IdentiTOP* in nicht ihrer Privatsphäre gestört zu fühlen. Um das zu überprüfen müsste eine Umfrage mit *IdentiTOP*-Nutzer:innen durchgeführt und ausgewertet werden. Das führt jedoch für dieses Beispiel zu weit. Daher lassen wir auch die Behandlung von R5 in diesem Beispiel aus. In Wirklichkeit wäre es aber natürlich sehr wichtig, R5 auch zu prüfen.

Unterrichtsidee: Fordern Sie die Lernenden auf zu überlegen, wie man die Risiken mit Hilfe der bereits bekannten Qualitätsmetriken überprüfen könnte. Sie können sich der Lösung gemeinsam nähern.

4.3. Qualitätsmetriken berechnen

An Hand der Modellergebnisse (Vorhersagen und Wahrheit) werden nun die ausgewählten Metriken berechnet.

Für die Fairness-Risiken werden die passenden Qualitätsmetriken für jede (intersektionale) Gruppe separat berechnet. Diese separate Qualitätsberechnung nennt man auch “Schnittanalyse” (en: slicing analysis).

Für den Vergleich der Qualitätsmetriken zweier Gruppen wird die Differenz oder das Verhältnis der Ergebnisse zueinander berechnet. Zum Vergleich der Qualität bei mehr als zwei Gruppen kann man zum Beispiel den Wert jeder Gruppe mit der am besten abschneidenden Gruppe vergleichen. Auch ein Vergleich mit dem durchschnittlichen Qualitätswert aller Gruppen wäre eine sinnvolle Variante.

Nicht jede Metrik lässt sich direkt an Hand der Modellergebnisse berechnen. Es müssen möglicherweise weitere Daten erhoben bzw. abgefragt werden. Um Fairness-Definitionen prüfen zu können, müssen z.B. demographische Daten vorhanden sein.

Demographische Daten für die Fairness-Prüfung von *IdentiTOP*:
Für R3 muss das Geschlecht und die Hautfarbe der Personen im Testdatenset bekannt sein. Um ein valides Ergebnis zu erhalten, wurde beim gezielten Test von IdentiTOP eine balancierte Auswahl an Testpersonen getroffen und ihre demographischen Daten als binärer Wert erhoben.

Es ergeben sich für die Personengruppen die folgenden Wahrheitsmatrizen:

Gruppe	**TP**	**FN**	**FP**	**TN**	∑
Männlich	56	9	5	80	150
Weiblich	57	18	5	70	150
Hellhäutig	64	3	4	77	148
Dunkelhäutig	49	24	6	73	152
Männlich und hellhäutig	30	2	3	39	74
Männlich und dunkelhäutig	26	7	2	41	76
Weiblich und hellhäutig	34	1	3	38	74
Weiblich und dunkelhäutig	23	17	4	32	76

Unterrichtsidee: Um das Verständnis der Wahrheitsmatrix zu üben können Sie den Lernenden Fragen zu den Daten stellen, z.B.: Wie viele hellhäutige Männer, die bei der Prüfung versucht haben zu betrügen, wurden von *IdentiTOP* erwischt?

Qualitätsberechnung für *IdentiTOP*:

- (R1) FNR ≈ 0,19
- (R2) FPR ≈ 0,06
- (R3 a)
  - Geschlecht
    - FNR(männlich) ≈ 0,14
    - FNR(weiblich) ≈ 0,24
    - FNR(w) / FNR(m) ≈ 1,73 > 1,2
      FNR(w) - FNR(m) ≈ 0,10
  - Hautfarbe
    - FNR(hellhäutig) ≈ 0,04
    - FNR(dunkelhäutig) ≈ 0,33
    - FNR(d) / FNR(h) ≈ 7,34 > 1,2
      FNR(d) - FNR(h) ≈ 0,28 > 0,2
  - Intersektionale Gruppen
    - FNR(männlich und hellhäutig) ≈ 0,06
    - FNR(männlich und dunkelhäutig) ≈ 0,21
    - FNR(weiblich und hellhäutig) ≈ 0,03
    - FNR(weiblich und dunkelhäutig) ≈ 0,43

        - Der beste FNR-Wert ist ~0,03 (für weiblich und hellhäutig). Der durchschnittliche FNR-Wert ist ~0,18. Mit diesen FNR-Werten wird nun jede verbliebene Gruppe verglichen:
            - FNR(mh) / FNR(min) ≈ 2,19 > 1,2
              FNR(mh) - FNR(min) ≈ 0,03
              FNR(mh) / FNR(durchschnitt) ≈ 0,34 < 0,8
              FNR(mh) - FNR(durchschnitt) ≈ -0,12
            - FNR(md) / FNR(min) ≈ 7,42 > 1,2
              FNR(md) - FNR(min) ≈ 0,18
              FNR(md) / FNR(durchschnitt) ≈ 1,17
              FNR(md) - FNR(durchschnitt) ≈ 0,03
            - FNR(wh) / FNR(durchschnitt) ≈ 0,16 < 0,8
              FNR(wh) - FNR(durchschnitt) ≈ -0,15
            - FNR(wd) / FNR(min) ≈ 14,88 > 1,2
              FNR(wd) - FNR(min) ≈ 0,40 > 0,2
              FNR(wd) / FNR(durchschnitt) ≈ 2,33 > 1,2
              FNR(wd) - FNR(durchschnitt) ≈ 0,24 > 0,2

- (R3 b)
    - Geschlecht
        - FPR(männlich) ≈ 0,06
        - FPR(weiblich) ≈ 0,07
        - FPR(w) / FPR(m) ≈ 1,13
          FPR(w) - FPR(m) ≈ 0,01
    - Hautfarbe
        - FPR(hellhäutig) ≈ 0,05
        - FPR(dunkelhäutig) ≈ 0,08
        - FPR(d) / FPR(h) ≈ 1,54 > 1,2
          FPR(d) - FPR(h) ≈ 0,03
    - Intersektionale Gruppen
        - FPR(männlich und hellhäutig) ≈ 0,07
        - FPR(männlich und dunkelhäutig) ≈ 0,05
        - FPR (weiblich und hellhäutig) ≈ 0,03
        - FPR(weiblich und dunkelhäutig) ≈ 0,11
        - Der beste FPR-Wert ist ~0,03 (für weiblich und hellhäutig). Der durchschnittliche FPR-Wert ist ~0,06. Mit diesen FPR-Werten wird nun jede verbliebene Gruppe verglichen.
            - FPR(mh) / FPR(min) ≈ 2,79 > 1,2
              FPR(mh) - FPR(min) ≈ 0,05
              FPR(mh) / FPR(durchschnitt) ≈ 1,12
              FPR(mh) - FPR(durchschnitt) ≈ 0,01
            - FPR(md) / FPR(min) ≈ 1,81 > 1,2
              FPR(md) - FPR(min) ≈ 0,02
              FPR(md) / FPR(durchschnitt) ≈ 0,73 < 0,8

FPR(md) - FPR(durchschnitt) ≈ -0,02
- FPR(wh) / FPR(durchschnitt) ≈ 0,40 < 0,8
  FPR(wh) - FPR(durchschnitt) ≈ -0,04
- FPR(wd) / FPR(min) ≈ 4,33 > 1,2
  FPR(wd) - FPR(min) ≈ 0,09 < 0,2
  FPR(wd) / FPR(durchschnitt) ≈ 1,75 > 1,2
  FPR(wd) - FPR(durchschnitt) ≈ 0,05

Unterrichtsidee: Berechnen Sie FOR und TPR gemeinsam und teilen Sie die Berechnung der Fairness-Definitionen auf. Die Berechnungen können auf Papier, per Exel oder mit einer beliebigen Programmiersprache - z.B. in Python mit dem beigefügten Jupyter Notebook - durchgeführt werden.

4.4. Qualitätsmetriken interpretieren

Als letzter Schritt werden die Berechnungen interpretiert. Lassen die Werte auf eine Erfüllung der Qualitätskriterien schließen? Welche Risikohypothesen haben sich bewahrheitet?

Ab welchem Schwellenwert man schließt, dass die Qualität ähnlich genug ist, um fair zu sein, muss je nach Kontext bestimmt werden. Häufig orientiert man sich an der “Disparate Impact”-Regel [12]. Danach sollte das Verhältnis zweier Qualitätsergebnisse zwischen 0,8 und 1,2 liegen, bzw. nach anderer Interpretation die Differenz zwischen -0,2 und 0,2 liegen. Wenn die Qualitätswerte sehr klein sind oder durch 0 geteilt werden müsste, bietet es sich an, die Differenz dem Verhältnis als Maß für Gleichheit vorzuziehen.

Häufig fällt die Aussage unterschiedlich aus, je nachdem welche Vergleichsart und welche Toleranzgrenzen gewählt wurden. Daher sollten stets alle Berechnungen offengelegt und ganzheitlich interpretiert werden.

Interpretation der Ergebnisse der Qualitätsberechnung für *IdentiTOP*:
- (R1) Von allen positiven Fällen (bei denen die Person nicht betrogen hat) wurden 20% fälschlicherweise als betrügend eingestuft. Das heißt, dass in einem von fünf Betrugs-Alarmen eine unschuldige Person in ihrem Prüfungsversuch gestört wird. Diese Risikohypothese hat sich bewahrheitet.
- (R2) 6% der betrügenden Prüflinge wurden vom Modell nicht bemerkt. Das heißt, dass das Risiko, dass sich Prüflinge unbehelligt Prüfungsleistungen erschleichen, eher unwahrscheinlich ist und darum akzeptabel ist.
- (R3 a) Bei nicht-betrügenden Personen werden häufiger dunkelhäutige Personen, insbesondere dunkelhäutige Frauen (doppelt so oft als dunkelhäutige Männer), fälschlicherweise als betrügend eingestuft als hellhäutige Personen. Hellhäutige Frauen erhalten am seltensten einen falschen Alarm. Hellhäutige Männer haben nur wenig mehr Alarme. Es liegt also keine allgemeine Diskriminierung nach

Geschlecht vor, sondern nach Hautfarbe, und darüber hinaus eine Diskriminierung dunkelhäutiger Frauen.
- (R3 b) Die FPR ist für alle Gruppen relativ gering, doch es lässt sich erkennen, dass das Modell dunkelhäutige Betrüger:innen etwas häufiger nicht bemerkt als hellhäutige Betrüger:innen. Betrügende hellhäutige Frauen werden am häufigsten enttarnt, dunkelhäutige Männer am zweithäufigsten, dann folgen hellhäutige Männer und schließlich dunkelhäutige Frauen, deren Betrug in 11% der Fälle nicht erkannt wird.
- Fazit: Es kann geschlossen werden, dass *IdentiTOP* Betrugsversuche, wenn sie vorkommen, im Allgemeinen auch als solche erkennt - allerdings werden dunkelhäufige Betrüger:innen häufiger nicht entdeckt als hellhäutige Betrüger:innen. Der Unterschied ist vor allem bei Frauen zu beobachten. Im Allgemeinen werden außerdem ein nicht zu vernachlässigender Anteil von unschuldigen Prüflingen in ihrem Prüfungsversuch gestört, was zu ungerechtfertigter Belastung von Lernenden führt. Bei dunkelhäutigen Personen - und besonders dunkelhäutigen Frauen - kommen wesentlich mehr falsche Alarme vor als bei hellhäutigen Personen. *IdentiTOP* benachteiligt also dunkelhäutige Personen und insbesondere dunkelhäutige Frauen.

Unterrichtsidee: Fordern Sie die Lernenden auf, die Ergebnisse zu interpretieren.

4.5. Verbesserungsvorschläge erarbeiten

Personen mit Vorwissen über Machine Learning können außerdem überlegen, was Gründe für die Ergebnisse sind (siehe auch [14]), und Empfehlungen abgeben, wie versucht werden könnte, das System zu verbessern.

Empfehlung:
Die hohe Belastung für unschuldige Prüflinge könnte durch eine “Human in the Loop”-Strategie vermieden werden, bei der ein Mensch im Fall eines Betrugsalarms die Identität der Person auf Grund der Fotos prüft, und den Prüfungsabbruch nur stattgibt, wenn der Mensch den Betrug bestätigen kann.
Um jedoch einem der ursprünglichen Ziele von *IdentiTOP*, Aufsichtskräfte zu entlasten, gerecht zu werden, sollte die Qualität des Modells weiter verbessert werden. Der Grund der hohen FNR und der FNR- und teilweise FPR-Ungleichheit könnte in den Trainingsdaten von *IdentiTOP* liegen. IdentiTOP sollte mit einem Trainingsdatensatz trainiert werden, der mehr dunkelhäutige Personen, insbesondere dunkelhäutige Frauen, enthält, und dann erneut getestet werden.

Unterrichtsidee: Fragen Sie die Lernenden, woher die Unfairness des Modells kommen könnte und was getan werden könnte, um die Risiken des Modells zu minimieren.

5. Zusammenfassung & Fazit

Machine Learning ermöglicht die Entwicklung von Modellen, die Daten klassifizieren und Vorhersagen treffen. Das Themengebiet "Faires Machine Learning" betrachtet die Qualität von Machine Learning Modellen nicht nur aus technischer Perspektive sondern auch aus sozialer Perspektive, ergründet Ursachen für Unfairness und sucht nach Lösungen. Es ist von entscheidender Bedeutung, die Fairness von KI-Systemen, insbesondere in sensiblen Bereichen wie der Gesichtserkennung, kontinuierlich zu überprüfen und zu verbessern, um Diskriminierung und Voreingenommenheit zu vermeiden.

Unterrichtsidee: Fordern Sie die Schüler:innen auf, in drei Sätzen schriftlich zu formulieren, welche Erkenntnisse sie über die Fairness von Machine Learning und die Bedeutung von Gesichtserkennungstechnologie aus dieser Unterrichtseinheit gewonnen haben.

Quellen

1) BesteTipps (2023). Gesichtserkennung auf Snapchat: wie geht das? Lösung. Verfügbar unter: https://www.bestetipps.de/computer/handy/gesichtserkennung-auf-snapchat-wie-geht-das-loesung/ (Zugriff zuletzt: 02.08.2023)
2) Vossen, R. (2013). Zukunft der Werbung: „Cara“ erkennt Alter und Geschlecht – und zeigt das passende Plakatmotiv. Verfügbar unter: https://www.basicthinking.de/blog/2013/05/22/zukunft-der-werbung-cara-erkennt-alter-und-geschlecht-und-zeigt-das-passende-plakatmotiv/ (Zugriff zuletzt: 02.08.2023)
3) adaLearning (2019). Joy Boulamwini on Face Recognition Technology. Verfügbar unter: https://www.youtube.com/watch?v=rWMLcNaWfe0 (Zugriff zuletzt: 07.08.2023)
4) Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of Machine Learning Research (Vol. 81, pp. 1-15). Conference on Fairness, Accountability, and Transparency. (Zugriff zuletzt: 07.08.2023)
5) Leisegang, D. (2020). Biometrische Videoüberwachung: Die neue Hochrisikotechnologie. Blätter für deutsche und internationale Politik. Verfügbar unter: https://www.blaetter.de/ausgabe/2020/maerz/biometrische-videoueberwachung-die-neue-hochrisikotechnologie (Zugriff zuletzt: 07.08.2023)
6) Stiebert, J. (2022). Klage gegen Gesichtserkennung bei Uni-Prüfungen. Posteo News. Verfügbar unter: https://posteo.de/news/klage-gegen-gesichtserkennung-bei-uni-pr%C3%BCfungen (Zugriff zuletzt: 07.08.2023)
7) Wang, N. (2019). Totale Überwachung in Chinas Schulen: Wenn Kameras jede Gesichtsregung auswerten. Der Tagesspiegel. Verfügbar unter: https://www.tagesspiegel.de/politik/wenn-kameras-jede-gesichtsregung-auswerten-4657504.html (Zugriff zuletzt: 07.08.2023)
8) Schiller, A. (2021). Universitäten spähen Studenten mit Software aus. Frankfurter Allgemeine Zeitung. Verfügbar unter:

https://www.faz.net/aktuell/karriere-hochschule/proctorio-und-wiseflow-hochschulen-spaehen-studenten-aus-17455837.html (Zugriff zuletzt: 07.08.2023)

9) Oppermann, A. (2021): Accuracy, Precision, Recall, F1-Score und Specificity. Verfügbar unter: https://artemoppermann.com/de/accuracy-precision-recall-f1-score-und-specificity/ (Zugriff zuletzt: 08.08.2023)
10) Hill, K., (2023): Eight Months Pregnant and Arrested After False Facial Recognition Match. Verfügbar unter: https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html (Zugriff zuletzt: 09.08.2023)
11) Wikipedia (2023): Confusion matrix. Verfügbar unter: https://en.wikipedia.org/wiki/Confusion_matrix (Zugriff zuletzt: 19.10.2023)
12) Wikipedia (2023): Disparate impact. Verfügbar unter: https://en.wikipedia.org/wiki/Disparate_impact (Zugriff zuletzt: 19.10.2023)
13) Verma, S. und Rubin, J. (2018). Fairness Definitions Explained. FairWare. Verfügbar unter: https://doi.org/10.1145/3194770.3194776
14) Suresh, H. und Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. EEAMO'21: Equity and Access in Algorithms, Mechanisms, and Optimization. Verfügbar unter: https://doi.org/10.1145/3465416.3483305