Technischer Appendix

Auswahl- und Einstellungsprozesse sind Ja/Nein Entscheidungen, welche sich methodisch durch klassische binäre Modelle analysieren lassen. Der *FAIR Index* zeigt diskriminierende Tendenzen auf unabhängig davon ob im Recruitingprozess die Auswahl mit oder ohne Algorithmus getroffen wurde.

Ein Algorithmus erhält Trainingsdaten $D = (x_i, y_i)_{i=1}^{n}$ bestehend aus *n* Instanzen, dem Feature-Vektor $x_i \in X$ für das Individuum *i* (x_i besteht beispielsweise aus Individuum *i*'s Alter, Geschlecht, Herkunft) und $y_i \in Y$, dem Ergebnis für Individuum *i* .
Im Folgenden wird angenommen, dass $X \subseteq \mathbb{R}^M$, *wobei M* der Anzahl der Features gleicht.
Da die Ergebnismenge *Y* im Auswahlprozess eine endliche Menge von Labels annimmt, nämlich *Y = {0,1}* handelt es sich um ein *binäres Classification Problem.* Entweder wurde das Individuum i eingestellt, $y_i = 1$, oder nicht $y_i = 0$.

Bei diskriminierungsbezogenen Fragestellungen werden Unterschiede in Bezug auf sensible Features untersucht. Eben solche spezifizieren die Zugehörigkeit eines Individuums *i* zu einer sozial hervorstechenden Gruppe wie Männer, Frauen, Ausländer oder Inländer. Die Menge an sensiblen Features für ein Individuum *i* wird durch $z_i \in Z = \{1,2,\ldots,K\}$ beschrieben. Ob z_i Bestandteil des Feature Vektors x_i ist, oder nicht, bleibt freigestellt. Die Partition der Trainingsdaten D, abhängig von den sensiblen Features, ergibt sich wie folgt:

$$D_z = \{(x_i, y_i) \mid z_i = z\}$$

D_z beschreibt die Gruppe eines sensiblen Features.

Im vorliegenden binären Classification Problem können alle Ergebnisse in einer Confusion Matrix *M* dargestellt werden. Diese ist durch die folgenden vier Zustände definiert:

(1) Richtig Positiv $(\hat{y} = 1, y = 1)$
(2) Richtig Negativ $(\hat{y} = 0, y = 0)$
(3) Falsch Positiv $(\hat{y} = 1, y = 0)$ und
(4) Falsch Negativ $(\hat{y} = 0, y = 1)$

Die Wahl der Optimierungsmaxime bestimmt die relative Wichtigkeit der verschiedenen Ergebniszuständen und legt verschiedenen Auffassungen von Fairness zu Grunde.

Im Rahmen von Auswahl- und Einstellungsprozessen ist die Chancengleichheit zwischen Gruppen ein häufig verwendetes Kriterium. So fordern wir im *FAIR Index* Setup eine Ausgeglichenheit der Falsch Positiven und Falsch Negativen über die betrachteten Gruppen des sensiblen Features. Ein Individuum *i* gilt als Falsch Positiv, wenn es eingestellt wurde, obwohl es laut neutralen eignungsdiagnostischen Verfahren (kognitiver Leistungstest oder CASE Score) geeignetere Kandidierende gab. Falsch Negativ definiert sich vice versa.

Die Differenz eines sensitiven Features zwischen den Falsch Positiven und Falsch Negativen Kandidierenden in Relation zu der Anzahl der Bewerbungen der sensitiven Gruppe, n_z, lässt sich wie folgt definieren:

$$Diff_z = \frac{D_z(\hat{y} = 1, y = 0) - D_z(\hat{y} = 0, y = 1)}{n_z}$$

Der *FAIR Index* ergibt sich über die Differenz der Ausprägungen des sensiblen Features z. Am Beispiel des Geschlechts, also z_i = {Männlich, Weiblich}, errechnet sich der *FAIR Index* folgendermaßen:

$$FAIRIndex = (Diff_{Maennlich} - Diff_{Weiblich}) * 100 ,$$

wobei

$$Diff_{Maennlich} = \frac{D_{Maennlich}(\hat{y} = 1, y = 0) - D_{Maennlich}(\hat{y} = 0, y = 1)}{n_{Maennlich}}$$

und

$$Diff_{Weiblich} = \frac{D_{Weiblich}(\hat{y} = 1, y = 0) - D_{Weiblich}(\hat{y} = 0, y = 1)}{n_{Weiblich}} \quad .$$

Der FAIR Index $F \in [-200, 200]$ für das Feature *Geschlecht* gibt nun an in wie weit Diskriminierung zwischen den Gruppen vorliegt (Abbildung 1). Ein FAIR Index von 0 bedeutet, dass die Anzahl der „fälschlicherweise“ eingestellten und nicht eingestellten Bewerber*innen über das sensible Feature hinweg gleich ist. Am Beispiel des biologischen Geschlechts würde dies bedeuten, dass genauso viele Männer wie Frauen den Job „unverdienterweise“ nicht bekommen haben, beziehungsweise „fälschlicherweise" eingestellt wurden.

Obwohl theoretisch Werte von (-)200 möglich sind, so beschreibt dies das extreme Szenario der maximalen Diskriminierung, in dem alle Individuen *i* welcher einer Gruppe des sensiblen Merkmals *z, also* D_z, angehören gemäß eignungsdiagnostischer Verfahren besser geeignet wären, jedoch alle verfügbaren Jobs durch die Zugehörige der entgegengesetzten Gruppe besetzt werden. In der Realität sind Personen bestimmter Gruppen nicht durchweg besser oder schlechter geeignet. Bei gleicher durchschnittlicher Eignung reduziert sich die Skala des FAIR Index auf $F \in [-100, 100]$.
Werte größer |100| sind somit in der praktischen Anwendung nicht zu erwarten. Deshalb sprechen wir bei einem FAIR-Indexwert von 100 bereits von einer maximalen Diskriminierung.