Woche 08 Praktische Anwendungsbeispiele: Entscheidungsbäume & Fairness

Skript

Erarbeitet von
Manh Khoi Duong

Lernziele

- RAPP als Anwendungsbeispiel für die Umsetzung von Responsible AI beschreiben können
- Erläutern können, wie aus informatischer Sicht eine Responsible AI angestrebt werden kann
- RAPP als Anwendungsfall von Entscheidungsbäumen nennen können
- Fairness in Machine Learning, insbesondere Statistical Parity als Fairness-Metrik, erläutern können

Projekt „KI für alle"
HeiCAD – Heine Center for Artificial Intelligence and Data Science

Inhalt

Einstieg

In diesem Video werden wir ein Forschungsprojekt besprechen, das sich auf die Entwicklung eines KI-Systems zur Vorhersage von akademischen Leistungen an Universitäten konzentriert. Dies wird als „Academic Performance Prediction“ oder kurz APP-System bezeichnet. Hier ist das Ziel, frühzeitig Studienabbrecher*innen zu identifizieren und so den präventiven Einsatz von individuellen Unterstützungsmaßnahmen zu ermöglichen.

Responsible Academic Performance Prediction (RAPP)

Im Projekt RAPP,

Quelle [1]

das ist kurz für „Responsible Academic Performance Prediction“, arbeiten Informatiker*innen und Sozialwissenschaftler*innen an der HHU gemeinsam daran, APP-Systeme auch sozial verträglich zu machen.

Das Ziel dieses Projekts ist es, eine gesellschaftlich akzeptable Nutzung von KI-Systemen durch Berücksichtigung ethischer Aspekte und der Wahrnehmung der Betroffenen – also z. B. Studierenden – durch das System zu schaffen.
Um dies aus der informatischen Sicht umzusetzen, konzentrieren wir uns auf „Explainable AI“ (also erklärbare KI) und Fairness.

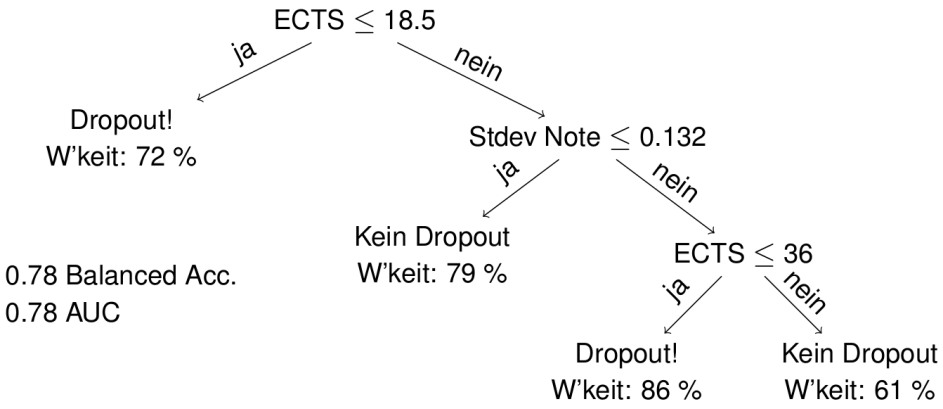
Explainable AI (XAI) ist die Fähigkeit von KI-Systemen, ihre Entscheidungen und Prozesse so zu erklären, dass sie von Menschen verstanden werden können. Für die Erklärbarkeit können Entscheidungsbäume als regelbasierte Erklärungskomponenten verwendet werden, um transparente Vorhersagen zu bekommen.

Zum Beispiel können Entscheidungsbäume verwendet werden, um die akademischen Leistungen von Studierenden anhand verschiedener Faktoren vorherzusagen. Diese Faktoren oder Features können z. B. Geschlecht, Alter, Nationalität, Zulassungsvoraussetzungen für die Universität, Anzahl bestandener Module und Anzahl erworbener Creditpoints im ersten Semester sein.
Akademische Leistungen, die vorhergesagt werden können, sind z. B.: Dropout-Risiko (bzw. die Wahrscheinlichkeit für den Abbruch des Studiums), Durchschnittsnote oder Studiendauer.

Beispiel für einen Entscheidungsbaum

Einblendung Grafik Entscheidungsbaum

Dropout via ECTS



Hier ist ein beispielhafter Entscheidungsbaum, der auf Daten zu akademischen Leistungen im 1. Semester von Studierenden der Informatik an der HHU trainiert wurde, um Dropouts – also Studienabbrüche – vorherzusagen. Wie man sieht, ist die Anzahl der gesammelten ECTS-Punkte für die Entscheidung, ob Studierende abbrechen, am relevantesten. Dies erkennt man daran, dass die ECTS-Punkte im Baum im obersten Knoten stehen. Hier gilt, dass man mit 72 % Wahrscheinlichkeit ein Dropout ist, wenn man weniger als 18.5 ECTS-Punkte im 1. Semester gesammelt hat. Sollte das nicht der Fall sein, verzweigt man hier im Baum nach rechts. Hier ist die Standardabweichung der Noten von Bedeutung. Bei einer niedrigeren Standardabweichung der Noten ist man mit 79 % Wahrscheinlichkeit kein Dropout. Das bedeutet: Sind sich die Noten der geschriebenen Klausuren ähnlicher, bricht man eher nicht ab. Das Konzept sollte nun klar sein: Um Vorhersagen treffen zu können, geht man entlang der Entscheidungen, bis man auf einem Blatt ankommt, worin die zugehörige Klasse für die Vorhersage steht. Wie man sieht, ist die Vorhersage für uns Menschen interpretierbar. Sollte die KI die Klasse durch soziodemografische Attribute vorhersagen, wäre solch eine Diskriminierung durch die KI direkt erkennbar. Explainable AI hilft uns also Entscheidungen der KI nachzuvollziehen, um mögliche Diskriminierung zu erkennen.

Fairness

Die Integration von Fairness in das APP-System ist ein wichtiger Aspekt, um sicherzustellen, dass die von dem System gemachten Vorhersagen nicht verzerrt sind und keine bestimmten Gruppen von Studierenden diskriminieren. Es gibt einige Möglichkeiten, wie Fairness in das System integriert werden kann.

Nun müssen wir uns die Frage stellen: Wie quantifiziert man jedoch Fairness im Kontext von Machine Learning?
Es gibt mehrere Fairness-Metriken (auch Fairness Notions genannt), die häufig verwendet werden, um die Fairness bzw. Diskriminierung eines Machine-Learning-Modells zu evaluieren.

Eine davon ist die *statistical parity*.

Quelle [2]

Statistical parity ist eine Fairness-Metrik, die misst, ob die Wahrscheinlichkeit von positiven Ergebnissen (wie Jobzusage, hohe Kreditwürdigkeit) für verschiedene Gruppen von Menschen gleich ist.
Für den APP-Anwendungsfall kann man z. B. die Wahrscheinlichkeit betrachten, den Kurs zu bestehen, unter der Bedingung man sei männlich/weiblich/nicht-binär. Hierbei sollen die Wahrscheinlichkeiten gleich sein. In der Praxis wird im Moment noch meistens zur Vereinfachung nur in binäre Gruppen wie männlich und weiblich unterteilt. Die Behandlung nicht-binärer Attribute für diverse Anwendungsfälle ist noch Gegenstand aktueller Forschung.

Um die Stärke der Diskriminierung zu messen, können wir z. B. den Unterschied zwischen der Wahrscheinlichkeit, den Kurs als männliche Person zu bestehen, und der Wahrscheinlichkeit, den Kurs als weibliche Person zu bestehen, anschauen. Je höher dieser Unterschied ist, desto stärker ist die Diskriminierung.

Wie schafft man es aber jetzt, unser Machine-Learning-Verfahren zur Vorhersage von Studienleistungen fair zu gestalten?
Es gibt eher naive Vorgehensweisen wie z. B. sicherzustellen, dass die zur Ausbildung des Systems verwendeten Daten vielfältig und repräsentativ für die Bevölkerung von Studierenden an der Universität sind. Dies kann dazu beitragen, verzerrte Vorhersagen aufgrund eines begrenzten oder unausgewogenen Datensatzes zu vermeiden.

Eine weitere Möglichkeit wäre es, Merkmale, die das System zur Vorhersage verwenden wird, sorgfältig auszuwählen. Einige Merkmale, wie Geschlecht oder Ethnizität, könnten möglicherweise zu verzerrten Vorhersagen führen, wenn sie nicht ordnungsgemäß berücksichtigt werden. Jedoch ist das Ausschließen nicht immer möglich, da man eine Proxy-Diskriminierung hervorrufen kann. Dies ist eine Diskriminierung, die durch eine Proxy-Variable passiert. Eine Proxy-Variable wird auch Stellvertreter-Variable genannt und gibt Auskunft über andere Variablen. Beispielsweise kann eine Proxy-Variable hierbei mit den soziodemografischen Attributen zusammenhängen und zu einer indirekten Diskriminierung führen. Ein Beispiel für eine Proxy-Variable wäre der Wohnort, anhand dessen man Aussagen über den Wohlstand machen kann.

Welche Methoden für die Gestaltung eines fairen Machine-Learning-Verfahrens gibt es?
Wir können an verschiedenen Stellen eines Machine-Learning-Verfahrens ansetzen, um es fair zu machen: beim Pre-Processing, also der Vorverarbeitung, beim In-Processing bzw. während des Trainings und beim Post-Processing bzw. Nachbearbeiten.

Beim Pre-Processing wird der gegebene Datensatz auf Bias untersucht. Diese werden anschließend entfernt, indem Datenpunkte gelöscht werden oder die Variablen im Datensatz so verändert werden, dass es fair ist. Zurückgegeben wird dann ein fairer Datensatz, mit dem Machine-Learning-Modelle trainiert werden können.

In-Processing-Verfahren modifizieren Machine-Learning-Algorithmen, indem Fairness beim Training berücksichtigt wird. Beispielsweise wird beim Training versucht, *statistical parity* zu erfüllen. Allgemein lernen Machine-Learning-Algorithmen, indem sie versuchen, den Fehler,

den sie bei der Vorhersage machen, zu minimieren. Meistens wird zusätzlich zum Fehlerterm ein Fairnessterm hinzugefügt, der die Diskriminierung misst. Dadurch wird nicht nur der Fehler, sondern auch die Diskriminierung wie beispielsweise die *statistical parity* Differenz minimiert.

Post-Processing-Verfahren korrigieren den Output von Vorhersagen, um fairer zu werden. So bekommen benachteiligte Gruppen statt negativer Vorhersagen vermehrt positive Vorhersagen, sodass die Differenz zu der privilegierten Gruppe minimal wird.

Schließlich ist es wichtig, die Fairness des Systems regelmäßig zu evaluieren, indem die von ihm gemachten Vorhersagen und die Ergebnisse für verschiedene Gruppen von Studierenden analysiert werden. Dies kann durch die Verwendung von Fairness-Metriken erfolgen, die die Fairness eines Systems anhand verschiedener Kriterien wie *statistical parity* sowie gleiche Vorhersagegenauigkeit und gleiche Fehlerraten messen. Bei den letzteren Kriterien möchte man verhindern, dass eine Gruppe mit einer höheren Fehlerquote Vorhersagen von der KI bekommt. Dazu gibt es auch Fallbeispiele aus der Medizin für People of Color.

Quelle [3]

Denn dies ist auch eine Diskriminierung. Insgesamt möchte man jedoch trotz Fairness nicht an der gesamten Vorhersagegenauigkeit schlechter werden. Dies ist in der Fachliteratur auch bekannt als Fairness-Accuracy Trade-off. Durch die regelmäßige Evaluierung der Fairness und Vorhersagegenauigkeit des Systems können mögliche Vorurteile identifiziert und behoben werden, sowie ein genaues System zum Einsatz gebracht werden.

Take-Home Message

Insgesamt erfordert die Integration von Fairness in das APP-System eine sorgfältige Überlegung und regelmäßige Evaluierung, um sicherzustellen, dass das System für alle Studierenden faire und unvoreingenommene Vorhersagen macht.

Quellen

Quelle [1] RAPP-Projekt-Homepage, https://rapp.hhu.de/

Quelle [2] Dunkelau, J., & Leuschel, M. (2019). Fairness-Aware Machine Learning: An Extensive Overview.

Quelle [3] https://www.aerzteblatt.de/archiv/224892/Haeufige-Dermatosen-Besonderheiten-bei-dunkler-Haut

Weiterführendes Material

Podcast InsideHeiCAD.
Staffel 2, Folge 6. https://www.heicad.hhu.de/aktivitaeten/der-heicadpodcast
Grah, J. (Moderatorin), Duong, M. K. (Gast). (2022, 9. November). #6: Wie quantifiziert man eigentlich Fairness? (PhD Pitches) [Audio-Podcast]. In *InsideHeiCAD*. Heine Center for Artificial Intelligence and Data Science.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214-226).

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. arXiv preprint arXiv:1609.07236.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017, April). Fairness constraints: Mechanisms for fair classification. In Artificial intelligence and statistics (pp. 962-970). PMLR.

Disclaimer

Transkript zu dem Video „Woche 08 Praktische Anwendungsbeispiele: Entscheidungsbäume & Fairness“, Manh Khoi Duong.