Datengenerator für Daten mit Bias als Grundlage für Data Science Projekte

Studienarbeit

für die Prüfung zum

Bachelor of Science

des Studiengangs Informatik

an der Dualen Hochschule Baden-Württemberg Stuttgart

von

Simon Jess, Timo Zaoral

Juni 2022

Bearbeitungszeitraum	04.10.2021 - 10.06.2022
Matrikelnummer, Kurs	8268544, 6146532, INF19C
Ausbildungsfirma	TRUMPF SE + Co. KG, Ditzingen
Betreuer	Prof. Dr. Monika Kochanowski

Erklärung

Wir versicherern hiermit, dass wir die vorliegende Studienarbeit mit dem Thema: *Datengenerator für Daten mit Bias als Grundlage für Data Science Projekte* selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt haben. Wir versichern zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Stuttgart, Juni 2022

Simon Jess

Timo Zaoral

Abstract

Fasst die Aufgabenstellung und Ergebnisse kompakt und übersichtlich in wenigen Zeilen zusammen (4-7 Zeilen).

Inhaltsverzeichnis

Abkürzungsverzeichnis

KI	Künstliche Intelligenz
bzw.	beziehungsweise

Abbildungsverzeichnis

Tabellenverzeichnis

1 | Einleitung

Die fortschreitende Digitalisierung ist kaum noch aus unserem Alltag wegzudenken. Durch immer mehr Programme, die einem den Alltag erleichtern sollen, nutzen wir die Errungenschaften der Digitalisierung täglich. Häufig ist hier die Rede von künstlicher Intelligenz. Dabei ist uns meist nicht einmal Bewusst, dass im Hintergrund mit künstlicher Intelligenz gearbeitet wird. Egal ob als intelligenten Routenplaner oder Sprachsteuerung, hinter all diese Anwendung steckt heute nicht mehr nur ein Optimierungsalgorithmus sondern Künstliche Intelligenz (KI).
Mit der Digitalisierung hat man begonnen große Datenmengen zu sammeln. Durch den technischen Fortschritt im Bereich von Big Data, werden diese Datenmengen heutzutage unvorstellbar groß. Mit dem Erfassen und Speichern von Daten ist man in der Lage seine Produkte stetig zu verbessern und sogar neue Geschäftsmodelle zu schaffen. Zu diesen neuen Geschäftsmodellen gehört auch die nicht mehr aus unserem Alltag wegzudenkenden KI. Sie ermöglicht es uns Entscheidungen zu treffen, wie sie auch ein Mensch treffen könnte, aber auch Vorhersagen zu machen, was zwar für den Menschen möglich ist, aber mit viel Aufwand verbunden ist. Egal ob eine Entscheidung oder eine Vorhersage von einer KI getroffen werden, dahinter stehen Daten die bereits gesammelt wurden und die Entscheidungsgrundlage für die KI bilden. Aus diesem Grund werden Daten eine wertvolle Ressource, sobald man anfängt die Daten zu verarbeitet und aktiv zu nutzen.
Für KI werden die Daten zum Lernen benutzt. Entscheidend für die Qualität der KI ist daher in den meisten Fällen die Datengrundlage auf der die KI besiert. Lernen bedeutet, dass Zusammenhänge und die dadurch abgebildeten Verhaltensweisen von der KI erkannt und sich selbst angeeignet werden. Durch diese Art des Lernens, wie auch wir Menschen unser Wissen erlernen, ergeben sich nicht nur Potentiale sondern auch Gefahren! Abhängig von der Datenqualität und Richtigkeit beziehungsweise (bzw.) Zuverlässigkeit der Daten werden zukünftige Entscheidungen und Vorhersagen getroffen. Eine KI betrachtet dabei die Daten vollkommen neutral ohne Hintergrundwissen über Richtigkeit und Zuverlässigkeit. Deshalb können Verzerrungen in den Daten durch die KI nicht erkannt werden. Diese Verzerrung wird auch Bias gennant und befindet sich in den Trainingsdaten mit denen die KI lernt. Die Folge daraus ist, dass sich KIs benachteiligende und diskriminierende Verhaltensweisen angeeignen und diese selbst in der Praxis ausüben.
Insbesondere für durch Computer getroffene Entscheidungen und Vorhersagen spielt die Ethik daher eine große Rolle. Diese kann in der Regel nicht aus den Daten erlernt werden und hängt von uns Nutzern ab. So stellt sich die Frage wie sollen Menschen mit Entscheidungen durch KI umgehen und sich auf diese Verlassen. Diese fehlende Ethik sorgt für

nicht zu Vernachlässigende Verzerrungen und bildet so einen Bestandteil der "Dark side of KI". beziehungsweise (bzw.)

1.1 Motivation

Eine KI und deren Entscheidungen basieren stets auf Daten aus der Vergangenhiet, den Trainingsdaten. Wenn diese Trainingsdaten durch einen Bias verzerrt sind, ist das nicht unbedingt bekannt. In den meisten Fällen ist eine solche Verzerrung verborgen und wird erst im produktiven Betrieb der KI festgestellt.

Diese Verzerrungen führen dann häufig zu Skandalen in der Medienwelt. Es wurde bereits diverse Male in der Presse darüber berichtet, dass bspw. in Unternehmen Bewerbungen durch ein KI vorsortiert werden und eine Diskriminierung in dem Muster der Auswahl erkennbar waren. Diese Diskriminierungen sind jedoch nicht zu vergessen immer auf Daten zurückzuführen und somit auch auf die Ersteller der Daten, also die Menschen dahinter.

- Anonymisierung/Pseudonymisierung bei besonders großen Datensätzen ist schwierig
- Nachvollziehbarkeit von Bias verzerrten Daten
- Veranschaulichung von Bias in Daten für die Allgemeinheit, um auf das Problem im Bereich ML aufmerksam zu machen

1.2 Zielsetzung

- Datengenerator für Bias verzerrte Daten
- Visualisierung von Bias in Lerndaten für ML
- Gesamt Produkt zur Erstellung von Daten und derer Bias Visualisierung für die Lehre

- 1 Satz, was sollen wir machen –> Stichwortliste mit Anforderungen

- Maschinelles Lernen hängt von den Trainingsdaten ab.
- Trainingsdaten können einen Bias Data enthalten.

1.3 Aufbau der Arbeit

2 | Stand der Technik

2.1 Daten

2.1.1 Datenqualität

2.1.2 Bias

- Begriffserklärung: Data Bias vs Bias Verzerrung (zu viel/zu wenig lernen im ml)
- Arten von Bias:
- Bias durch Abwesenheit - Wenn eine Info fehlt, kann das zu Diskriminierung führen.
- Diskriminierung durch Menschen.
Arten von Bias: Cognitive, Social, Perceptual und Motivational Bias [1]

2.2 Künstliche Intelligenz & Maschinelles Lernen

2.2.1 Was ist Künstliche Intelligenz

2.2.2 Teilgebiet Maschinelles Lernen

- Superviced learning
- Unsuperviced learning

2.2.3 Ethik in der Künstlichen Intelligenz

2.3 Künstliche Intelligenz verbunden mit Bias

2.3.1 Diskriminierung durch verzerrte Daten

2.3.2 Gegenmaßnahmen

- Wenn der Parameter mit dem Bias entfernt wird, wird das Ergebnis erstmal schlechter.

3 | Praktischer Teil

In diesem Teil der Arbeit werden zuerst die beiden Szenarien erläutert und daraufhin die Konzeption und Umsetzung derer in Python beschrieben.

3.1 Szenarien

Für das generieren von Daten wurden zwei möglichst reale Szenarien ausgewählt. Zum einen das Szenario eines Bewährungsantrages, für welches 5 verschiedene Attribute und eine endgültige Bewertung mit stattgegeben oder nicht generiert werden. Zum anderen das zweite Szenario des social creditpoint system, für welches pro Person 7 Attribute zu generieren sind und eine numerische Bewertung zwischen 600 und 1400 creditpoints erstellt wird. Diese beiden Szenarien werden im folgenden genauer erläutert.

3.1.1 Szenario 1

In Szenario 1 soll ein Bewährungsantrag einer Person Bewertet werden. Ein Antrag besteht dabei aus dem Namen der Person, dessen Geschlecht, Hautfarbe und den entscheidenden Attributen der laufenden Strafe in Jahre und der Härte des Vergehens. Basierend auf diesen Attributen soll ein Bewerter beurteilen, ob der Antrag genehmigt oder abgelehnt wird. Das Geschlecht wird in „Männlich“ und „Weiblich“ angegeben. Die Hautfarbe der Person wird als „Schwarz“ oder „Weiß“ festgehalten. Die noch laufende Strafe des Gefangenen wird in Jahren von als Ganzzahlen von 1-5 angegeben. Da hier definiert wird ein Bewährungsantrag kann erst ab maximal 5 Jahren noch offene Strafe gestellt werden. Die Härte des Vergehens wird einfachheitshalber in den Gruppen „Leicht“, „Mittel“ oder „Hart“ festgehalten.
Für die Beurteilung des Antrags von dem Bewerter werden folgende Regeln definiert:

Attribut	**Positive Auswirkung**	**Negative Auswirkung**
Laufende Strafe	1-3	4-5
Härte des Vergehens	Leicht, Mittel	Hart

Tabelle 3.1: Tabelle für die Auswirkung der Attributen von Szenario 1

Das Geschlecht und die Hautfarbe werden hierbei nicht direkt aufgelistet, da diese in der Regel keine Auswirkung auf die Bewertung haben sollten. Diese können jedoch durch

einen konkreten Bias Aussagekraft bekommen. Damit soll in den generierten Daten die gewünschte Verzerrung auf einen gewissen Wert gelegt werden können. In diesem Szenario sind die möglichen Werte, welche durch eine Verzerrung und damit einem menschlichem Vorurteil eines Bewerters beeinflusst werden können, das Geschlecht und die Hautfarbe. Die anderen beiden Attribute, welche in der Tabelle 3.1 aufgeführt sind, wirken sich durch ihre Ausprägungen positiv oder negativ auf die Bewertung des Antrages aus. So wirkt z.B. eine Härte des Vergehens vom Niveau Leicht sich eher für eine positive Bewertung des Antrages aus, als eine mittlere Härte. Dasselbe gilt auch für die Laufende Strafe. So kann ein Bewerter dann anhand dieser beiden Werte eine Tendenz erhalten und dann über die Gestattung des Antrages entscheiden.

3.1.2 Szenario 2

Im zweiten Szenario wird das durch China populär gewordene sozial creditpoint System in einer lagenunabhängigen Version nachgebaut. Dafür werden Einträge zu Personen erstellt, nach welchen die Punktzahl der einzelnen Person zwischen 600 und 1400 Punkten bestimmt wird. Ein Eintrag zu einer Person beinhaltet die sieben in der folgenden Tabelle dargestellten Attribute mit den unterschiedlichen Ausprägungen. Die in Tabelle 3.2

Attribut	**Ausprägungen**
Name	Beliebig
Alter	20-79
Politische Orientierung	Links, Mitte, Rechts
Bildungsabschluss	Ausbildung, Fachschulabschluss, Bachelor, Master, Diplom, Promotion, ohne
Soziales	0-3
Wohnlage	Großstadt, Kleinstadt, Vorort, Ländlich
CO2-Fußabdruck	4-12

Tabelle 3.2: Tabelle der Attribute und Auswirkungen von Szenario 2

aufgeführten Ausprägungen haben ähnlich wie zu Szenario 1 unterschiedlich starke Auswirkungen auf den am Ende bestimmten social Score. Einzig allein der Name und das Alter sollen keine direkte Auswirkung auf den social Score haben. Die anderen Attribute wirken sich je nach Auswirkung positiv durch eine Erhöhung des Scores oder negativ durch eine Verringerung des Scores aus. Insgesamt werden so in diesem Szenario viele Einträge von Personen erstellt, welche alle unterschiedlichste Verteilungen der Ausprägungen besitzen und dadurch in der Bewertung einen individuellen social Score erzielen. Um nun eine gewünschte Verzerrung in die Daten zu bekommen können alle Attribute bis

auf den Namen, welcher rein als Füllwert dient, durch eine Verzerrung beeinflusst werden. So können z.B. Personen zwischen 20-30 Jahre negativ verzerrt werden, da ein oder zwei Bewerter etwas gegen junge Leute haben und diesen aus ihrer Überzeugung einen schlechteren Score geben. In diesem Szenario ist somit eine hohe Variabilität geboten inwieweit eine Verzerrung in die Daten gebracht wird. Zudem kann auch eine Verzerrung über mehrere Attribute eingebracht werden, da ein Bewerter z.B. auch etwas gegen eine Rechte Politische Orientierung und ein schlechtes Soziales Engagement von 0 haben kann.

3.2 Konzeption

In diesem Kapitel wird die erarbeitete Konzeption für die Umsetzung der beiden im Kapitel 3.1 aufgeführten Szenarien erläutert. Dabei wird in ein Grobkonzept zur allgemeinen Generierung der Daten und darauf in ein Feinkonzept für jedes Szenario unterteilt.

3.2.1 Grobkonzept

Das Grobkonzept beinhaltet die Überlegungen, wie die Programme/Notebooks für die beiden Szenarien generell aufgebaut sein sollen. Der Ablauf der Programme von der Eingabe der Parameter bis hin zu den fertig generierten Daten wird in fünf Schritten durchgeführt. Der Ablauf der Schritte ist in folgendem Programmablaufplan dargestellt.

Die in der Abbildung 3.1 dargestellten Hauptschritte des Programmablaufs lauten: Parametereingabe, Generieren der Daten, Regeln aufstellen, Bewerten und Speichern der Daten. Im ersten Schritt Parametereingabe wird dem Benutzer die Möglichkeit gegeben die Parameter für die Generierung der Daten einzugeben, wie z.B. die Anzahl der Daten oder Bewerter welche generiert werden sollen. Im folge Schritt werden daraufhin die passende Anzahl an Daten für das jeweilige Szenario generiert. Dabei sollen die Daten möglichst an Verhältnissen aus der Realität angepasst und auf dieser Grundlage generiert werden.

3.2.2 Feinkonzept

3.3 Umsetzung

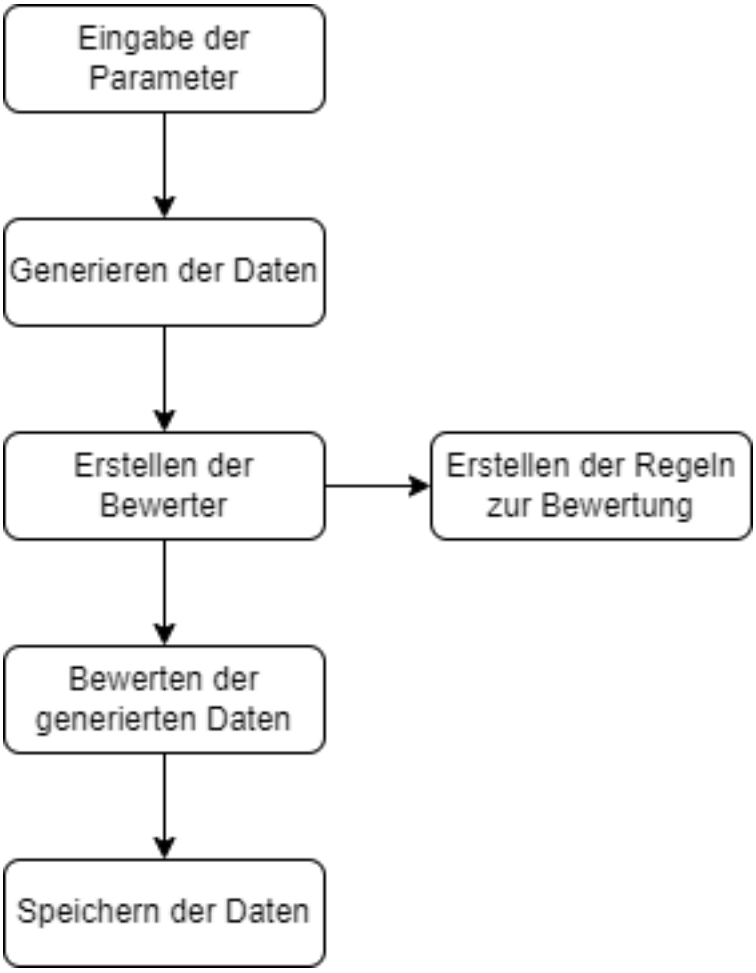

Abbildung 3.1: Ablauf der fünf Hauptschritte

4 | Schluss

4.1 Zusammenfassung

- Fazit ziehen!!!

4.2 Evaluierung

4.3 Ausblick

Literaturverzeichnis

[1] P. A., A. Jawaid, S. Dev, and V. M.S., "The patterns that don't exist : Study on the effects of psychological human biases in data analysis and decision making," in *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, pp. 193–197, 2018.