Vorwort

Willkommen beim Data-Science-Crashkurs. Wenn Sie bereits öfter etwas von Big Data, maschinellem Lernen, der künstlichen Intelligenz oder Data Science gehört haben und wissen wollen, welche Methoden sich hinter diesen Begriffen verbergen, sind Sie hier genau richtig. Dieses Buch richtet sich an alle, die mehr über die Möglichkeiten der Datenanalyse lernen wollen, ohne gleich tief in die Theorie oder bestimmte Methoden einzusteigen. Auch wenn Sie sich am besten schon etwas in der Informatik und/oder Mathematik auskennen, so kann man den Großteil auch verstehen, wenn man sich einfach nur für Daten interessiert und vor Mathe in der Schule keine Angst hatte.

Wir fangen an, indem wir die Definitionen der Begriffe einführen und dann betrachten, wie Data-Science-Projekte üblicherweise ablaufen. Dann geht's los mit den Daten. Zuerst lernen wir die Daten selbst mithilfe von Statistiken und Visualisierungen kennen. Sodann tauchen wir in die Welt der Algorithmik ein: Assoziationsregeln zum Auffinden von Beziehungen, Clustering, um Gruppen ähnlicher Daten zu finden, Klassifikation, um Kategorien zuzuweisen, Regression, um Zusammenhänge zu lernen, Zeitreihenanalyse, um zeitliche Zusammenhänge auszunutzen. Im letzten Teil betrachten wir, wie wir Texte zu Zahlen werden lassen, mit denen wir rechnen können, welche Rolle die Statistik für die Bewertung von Ergebnissen spielt und wie man mit Big Data arbeiten kann.

Wenn Sie wollen, können Sie sich nur mit den Methoden beschäftigen. Alle Methoden werden aber auch praktisch demonstriert: Im ganzen Buch befinden sich Quelltextbeispiele und das Ergebnis der Ausgabe. Hierdurch können Sie gleichzeitig verstehen, welche Methoden es gibt und wie Sie diese anwenden. Das Buch selbst wurde komplett mithilfe von *Jupyter Notebooks* geschrieben. In diesen Notebooks können Sie direkt im Webbrowser den Quelltext selbst ausführen, um die Ergebnisse interaktiv nachzuvollziehen. Sie können den Quelltext hierbei auch beliebig anpassen, zum Beispiel, um besser zu verstehen, was passiert, wenn sich Parameter eines Modells ändern. Im Anhang wird erklärt, wie Sie sich das Buch auf Ihrem eigenen Rechner »installieren«.

Am Ende der meisten Kapitel gibt es praktische Übungen, die Sie zur Vertiefung und für ein besseres Verständnis bearbeiten können. Beispiellösungen für diese Aufgaben können Sie sich in der Onlineversion des Buches anschauen[1]. Wenn Sie mal unterwegs sind, können Sie auch einfach in der Onlineversion weiterlesen. Für einige Kapitel gibt es keine Übungen. In diesen Kapiteln geht es großteils um Definitionen (Kap. 1 und 2) oder einen Ausblick (Kap. 13). Bei Kapitel 12 wird eine Big-Data-Umgebung benötigt. Auch wenn man hierfür kleinere Beispiele definieren könnte, müssten sowohl eine Python- als auch eine Java-Umgebung für eine Übung eingerichtet und passend konfiguriert werden, was aber über den bei einem Crashkurs angebrachten Aufwand hinausgeht.

Mit Quellenangaben wird in diesem Buch insgesamt eher sparsam umgegangen: Das Ziel ist ein breiter Überblick und ein Verständnis des Themas. Für die Vertiefung gibt es zu jedem Kapitel, häufig sogar zu den Abschnitten innerhalb der Kapitel, ausreichend eigene Fachliteratur. Welche Bücher jeweils für Sie geeignet sind, kann man nicht pauschal sagen, das hängt vom jeweiligen Ziel und Wissensstand der Leserschaft ab. Durch die aktuelle Verbreitung findet man zu jedem Thema auch zusätzliche Informationen im Internet, wenn man nach den entsprechenden Begriffen sucht. Die meisten Quellen, die hier im Buch genannt sind, verweisen auf besonders wichtige Definitionen oder Anwendungen; in wenigen Fällen, wenn ein Thema wirklich nur sehr kurz behandelt wird, auch zur weiterführenden Fachliteratur. In der Printversion des Buches sind diese Quellen durch die Autoren, Titel und Jahreszahlen angegeben, wie es in Literaturlisten üblich ist. In der Onlineversion wird direkt auf die Quellen verlinkt, falls möglich mithilfe von *Document Object Identifiern* (DOIs): Dies sind persistente Identifier, die auch in vielen Jahren noch funktionieren sollten und zu den Homepages der Verlage weiterleiten.

Der Fokus des Buches liegt darauf, wie Analysen für Daten erstellt werden. Drei für die Anwendung wichtige Aspekte betrachten wir nicht: Wie man Daten sammelt, wie man Daten aus einer Datenbank laden kann und den operativen Einsatz von erstellten Analysen. Das Datensammeln ist von Anwendungsfall zu Anwendungsfall verschieden. Oft liegen schon Daten vor. Andernfalls muss individuell auf Basis der Problemstellung eine Lösung entwickelt werden. Das Laden von Daten hängt vom Datenformat ab. Bei kleineren Projekten werden häufig CSV-Dateien verwendet, wie wir es in den Übungen machen. Aber auch andere Formate wie JSON sind bei Dateien üblich. Hier muss man einfach bei den Bibliotheken nach den entsprechenden Möglichkeiten zum Laden suchen. Im Fall von Datenbanken kann man die Daten häufig mit der Anfragesprache SQL laden. SQL ist für sich genommen jedoch bereits ein Thema, das ganze Bücher füllt. Da das Laden von Daten für unseren Crashkurs sekundär ist, verzichten wir daher auf

1. *https://data-science-crashkurs.de*

eine Einführung in Datenbanken und SQL. Beim operativen Einsatz geht es nicht nur um die Modelle und ihre Güte, sondern auch darum, wie diese in ein Softwaresystem eingebunden werden, wie dieses System getestet wird und wie hiermit bei einer Continuous Integration umgegangen wird. Hierbei handelt es sich um für die Softwaretechnik wesentliche Fragen, die jedoch die Analysen nicht direkt beeinflussen.

In diesem Buch verwenden wir vorwiegend die weibliche Bezeichnung von Rollen. Die Ausnahme ist die Rolle des Data Scientist. Im Englischen ist das Wort geschlechtslos und die Begriffe »Scientistin« oder »die Data Scientist« klingen eher komisch.

Ich wünsche Ihnen, liebe Leserinnen und Leser, viel Spaß und einen maximalen Erkenntnisgewinn mit diesem Buch.

Über Ihr Feedback würde ich mich freuen (*steffen.herbold@tu-clausthal.de*).

Steffen Herbold
Clausthal-Zellerfeld, Oktober 2021