Vorgehensmodelle für prädiktive Analysen

Dorian Zwanzig

About me

Dorian Zwanzig

M.Sc. Wirtschaftsingenieurwesen (FH)

Production Engineer

Bundesdruckerei GmbH

Production Industrial Engineering Data Analytics

Inhalte

Part 1

Zurücklehnen, Teetrinken & Zuhören

- Vorgehensmodelle im Überblick
- CRISP DM
- ML-Prozess nach Raschka

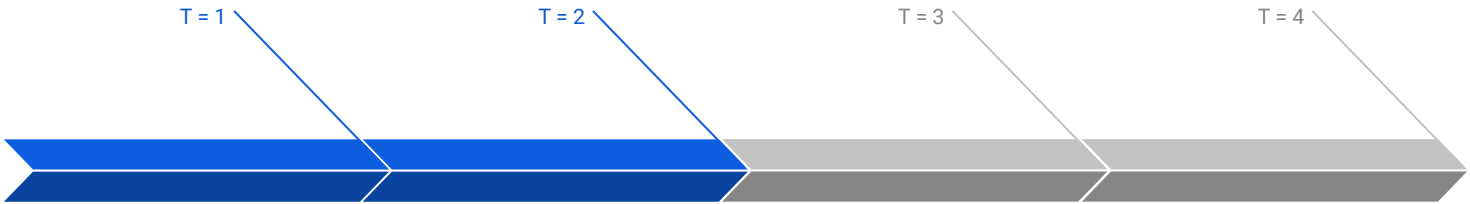
Part 2

Hands on

- Jupyter Notebooks 101
- Titanic Dataset

Part 1

Vorgehensmodelle
für prädiktive Analysen



Phase 1

Aufgaben A, B, C

Phase 2

Aufgaben A, B, C

Phase 3

Aufgaben A, B, C

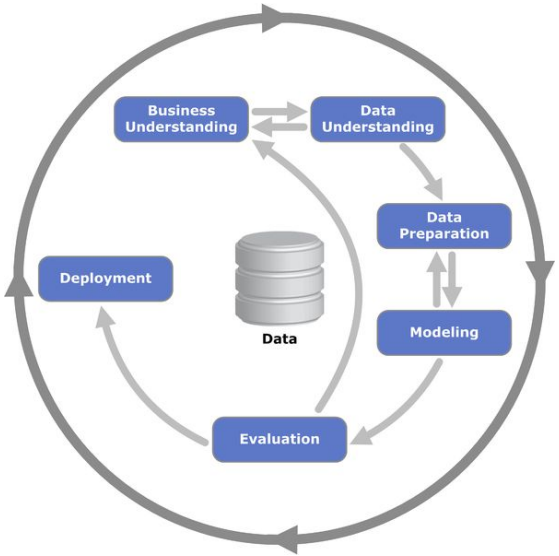
Phase 4

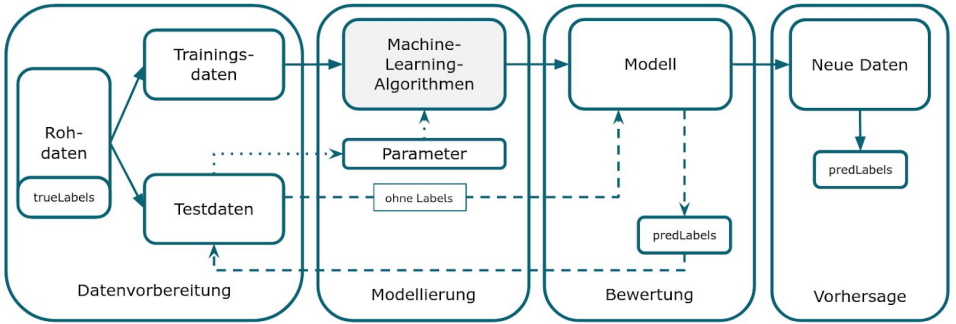
Aufgaben A, B, C

Vorgehensmodelle für **prädiktive Analysen**

Überblick Vorgehensmodelle

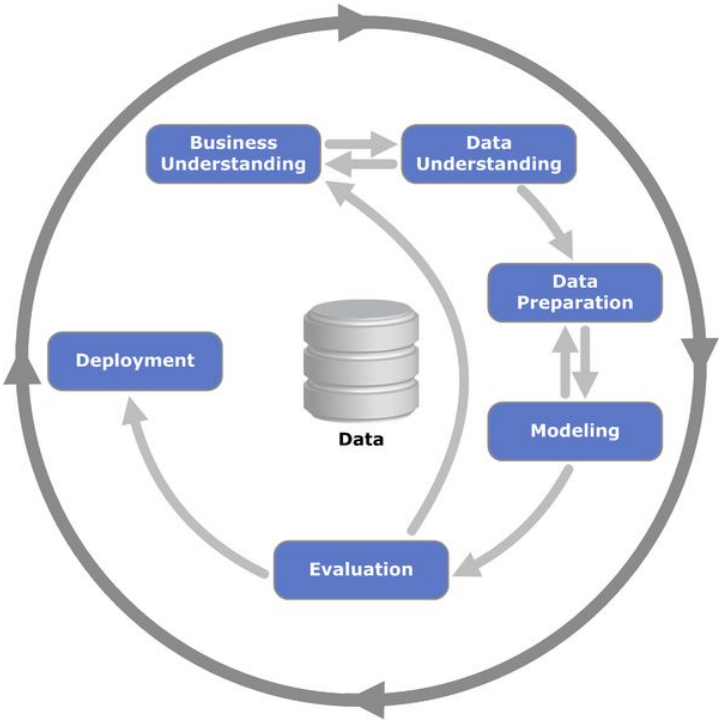
CRISP DM & ML nach Raschka





Quelle:
https://en.wikipedia.org/wiki/File:CRISP-DM_Process_Diagram.png
i.A.a. Raschka, Mirjalili (2018) Machine Learning with Python, mitp Verlag, 2. Auflage

CRISP DM Phasen & Vorgehen



Quellen:
https://en.wikipedia.org/wiki/File:CRISP-DM_Process_Diagram.png
https://www.bigdata-insider.de/was-ist-crisp-dm-a-815478/

CRISP DM Geschäftsverständnis

Festlegung der Ziele und Anforderungen;
Ableitung der konkreten Aufgabenstellung und der groben Vorgehensweise

Bestimmen der Geschäftsziele

- Erarbeiten des Geschäftshintergrunds
- Definieren von Geschäftszielen
- Kriterien für den Unternehmenserfolg

Bewerten der Situation

- Ressourcenbestand
- Anforderungen, Annahmen und Beschränkungen
- Risiken und Notfälle, Terminologie, Kosten-/Nutzen-Analyse

Quellen:
https://statistik-dresden.de/archives/1128
IBM SPSS ModelerCRISP-DM-Handbuch

CRISP DM Geschäftsverständnis

Festlegung der Ziele und Anforderungen;
Ableitung der konkreten Aufgabenstellung und der groben Vorgehensweise

Bestimmen von Data-Mining-Zielen

- Data-Mining-Ziele formulieren
- Erfolgskriterien für das Data-Mining festlegen

Erstellen eines Projektplans

Quellen:
https://statistik-dresden.de/archives/1128
IBM SPSS ModelerCRISP-DM-Handbuch

CRISP DM Datenverständnis

Datensammlung bzw. erste Sichtung der zur Verfügung stehenden Daten;
Ermittlung möglicher Probleme mit Datenqualität

Sammeln von Anfangsdaten

- Vorhandene Daten
- Erworbene Daten
- Zusätzliche Daten

Beschreiben von Daten

- Menge an Daten
- Werttypen
- Codierungsschemata

Quellen:
https://statistik-dresden.de/archives/1128
IBM SPSS ModelerCRISP-DM-Handbuch

CRISP DM Datenverständnis

Datensammlung bzw. erste Sichtung der zur Verfügung stehenden Daten;
Ermittlung möglicher Probleme mit Datenqualität

Untersuchen von Daten

- Explorative Datenanalyse
- Schreiben eines Berichts zur Datenexploration

Überprüfen der Datenqualität

- Qualitätative Datenanalyse
- Schreiben eines Berichts zur Datenqualtiät

Quellen:
https://statistik-dresden.de/archives/1128
IBM SPSS ModelerCRISP-DM-Handbuch

CRISP DM Datenvorbereitung

Konstruktion des finalen Datensatzes für die Modellierung

Auswählen von Daten

- Auswahl von Elementen (Zeilen)
- Auswahl von Attributen oder Merkmalen (Spalten)

Bereinigen von Daten

- Fehlende Daten
- Datenfehler
- Codierungsinkonsistenzen
- Fehlende oder ungültige Metadaten

Quellen:
https://statistik-dresden.de/archives/1128
IBM SPSS ModelerCRISP-DM-Handbuch

CRISP DM Datenvorbereitung

Konstruktion des finalen Datensatzes für die Modellierung

Erstellen neuer Daten

- Ableiten von Attributen (Spalten oder Merkmale)
- Generieren von Datensätzen (Zeilen)

Verbinden von Daten

- Verbinden von Daten
- Anhängen von Daten

Quellen:
https://statistik-dresden.de/archives/1128
IBM SPSS ModelerCRISP-DM-Handbuch

CRISP DM Modellierung

Anwendung geeigneter Data Mining-Verfahren, Optimierung der Parameter; gewöhnlich Ermittlung mehrerer Modelle

Auswählen der Modellierungsverfahren

- Berücksichtigung der verfügbaren Datentypen
- Berücksichtigung der Data-Mining-Ziele
- Bestimmen der Modellierungsanforderungen

Generieren eines Testdesigns

- Beschreiben der Kriterien für die "Güte" eines Modells
- Definieren der Daten, an denen diese Kriterien getestet werden

Quellen:
https://statistik-dresden.de/archives/1128
IBM SPSS ModelerCRISP-DM-Handbuch

CRISP DM Modellierung

Anwendung geeigneter Data Mining-Verfahren, Optimierung der Parameter; gewöhnlich Ermittlung mehrerer Modelle

Erstellen der Modelle

- Parametereinstellungen
- Ausführen der Modelle
- Modellbeschreibung

Modellbewertung

- Anwendung des Testdesigns
- Dokumentation der Testergebnisse

Quellen:
https://statistik-dresden.de/archives/1128
IBM SPSS ModelerCRISP-DM-Handbuch

CRISP DM Evaluierung

Auswahl des Modells, das die Aufgabenstellung am besten erfüllt.
Sorgfältiger Abgleich mit der Aufgabenstellung.

Evaluieren der Ergebnisse

Überprüfungsprozess

Bestimmen der nächsten Schritte

Quellen:
https://statistik-dresden.de/archives/1128
IBM SPSS ModelerCRISP-DM-Handbuch

CRISP DM Bereitstellung

Aufbereitung und Präsentation der Ergebnisse;
evtl. Integration des Modells in einen Entscheidungsprozess des Auftraggebers

Planen der Bereitstellung

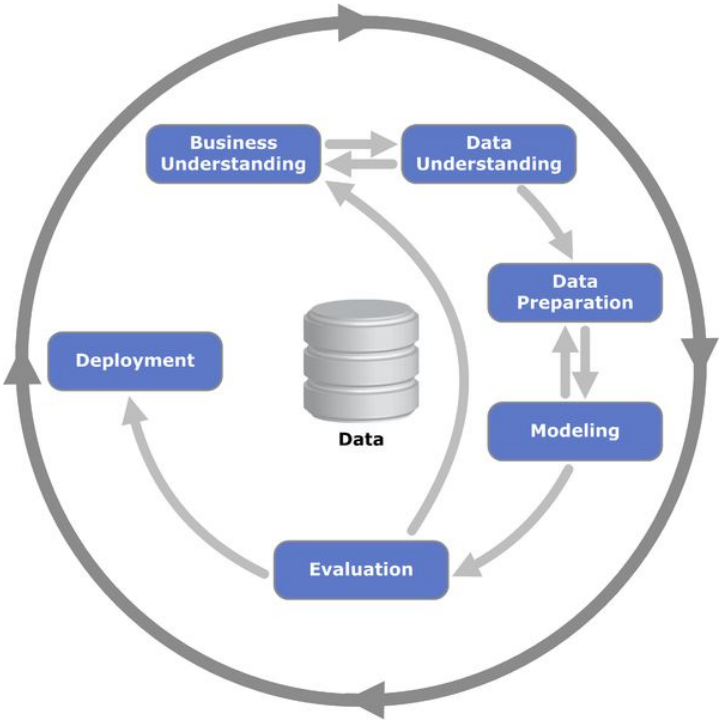
Planen von Überwachung und Anpassung

Erstellen eines Abschlussberichts

Durchführen einer abschließenden Projektbewertung

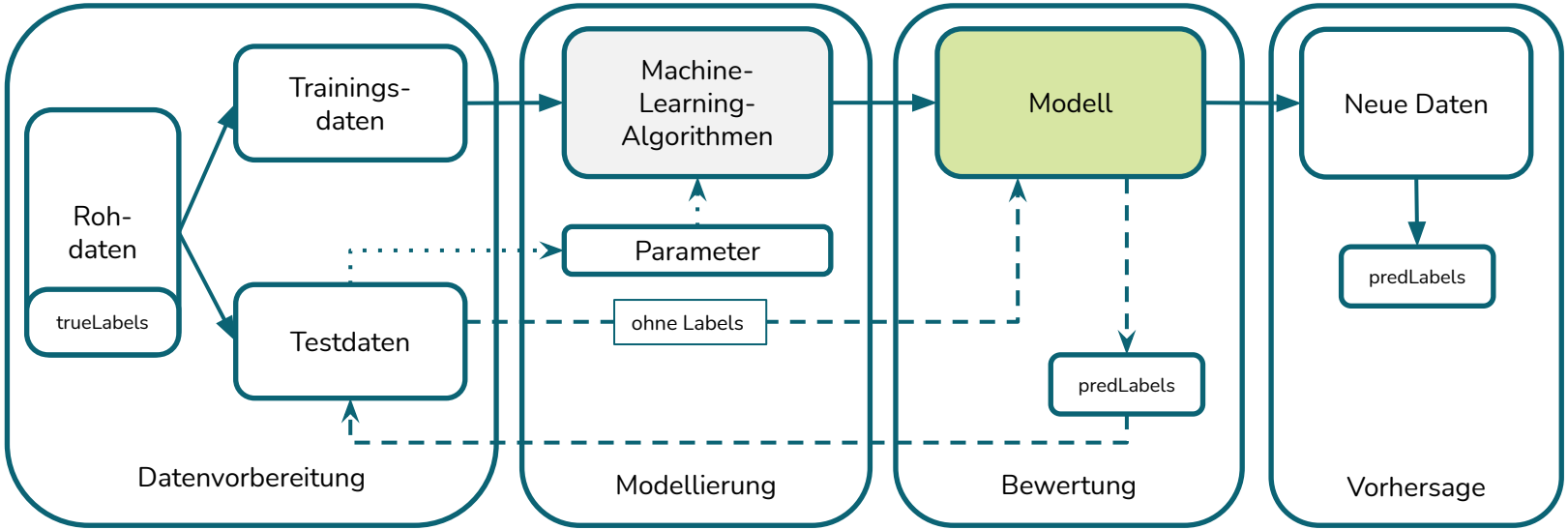
Quellen:
https://statistik-dresden.de/archives/1128
IBM SPSS ModelerCRISP-DM-Handbuch

CRISP DM Phasen & Vorgehen



Quelle:
https://en.wikipedia.org/wiki/File:CRISP-DM_Process_Diagram.png

ML-Prozess i.A.a. Raschka Überblick



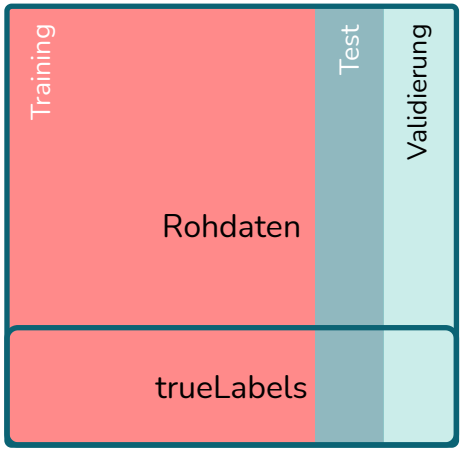
Quelle:
i.A.a. Raschka, Mirjalili (2018) Machine Learning with Python, mitp Verlag, 2. Auflage

ML-Prozess i.A.a. Raschka
Datenvorbereitung

Daten sind i.d.R. bereits bereinigt

Aufteilung des vollständigen Datensatzes in

- Trainingsdatensatz
- Testdatensatz
- opt. Validierungsdatensatz



Quelle:
i.A.a. Raschka, Mirjalili (2018) Machine Learning with Python, mitp Verlag, 2. Auflage

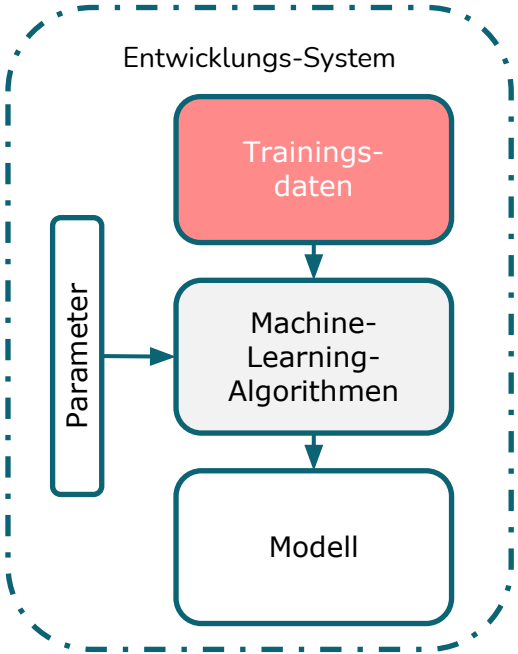
ML-Prozess i.A.a. Raschka
Modellierung

Auswahl und Initialisierung des Modells/ der Modelle

Nutzung des Trainingsdatensatzes zur Modellbildung (aka Anwendung ML-Algos)

Auswertung der Modellgüte anhand des opt. Validierungsdatensatzes

opt. Anpassung der Modell-Parameter (aka. Hyperparameter)



Quelle:
i.A.a. Raschka, Mirjalili (2018) Machine Learning with Python, mitp Verlag, 2. Auflage

ML-Prozess i.A.a. Raschka
Bewertung

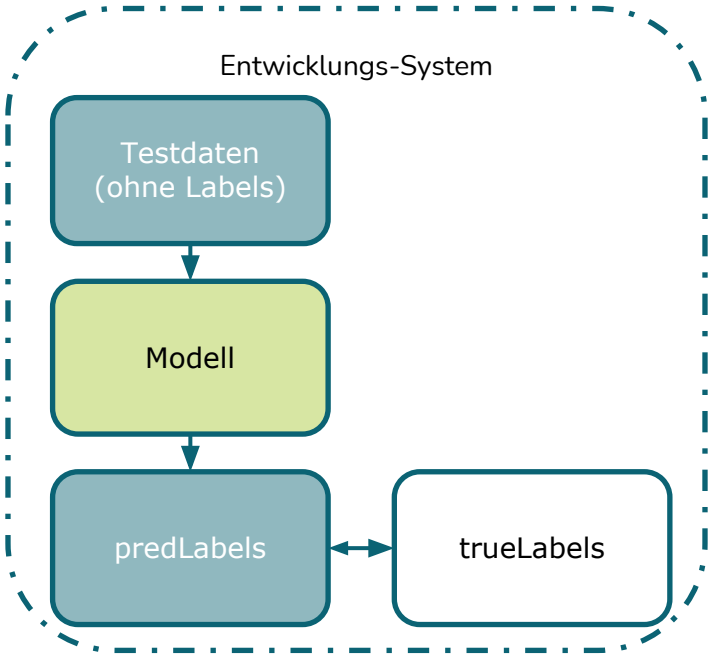
Auswahl der Bewertungsmethode

Anwendung des Modell auf Test-Datensatz

Bewertung des Modells

Modellvergleich

opt. Parameteranpassung und erneute Modellbildung

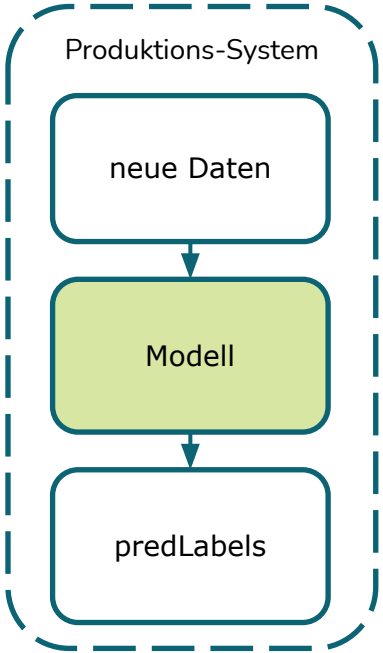


Quelle:
i.A.a. Raschka, Mirjalili (2018) Machine Learning with Python, mitp Verlag, 2. Auflage

ML-Prozess i.A.a. Raschka
Vorhersage

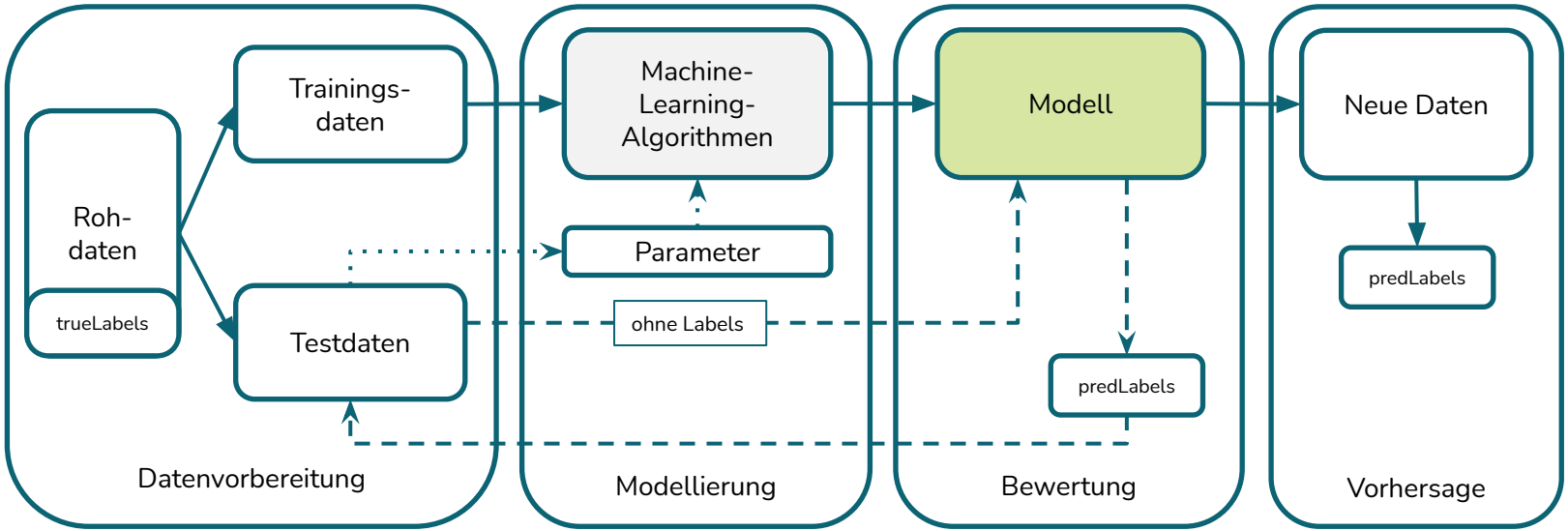
Implementierung des Modells in Anwendungsumgebung

Anwendung des Modell auf “unbekannte” Datensätze



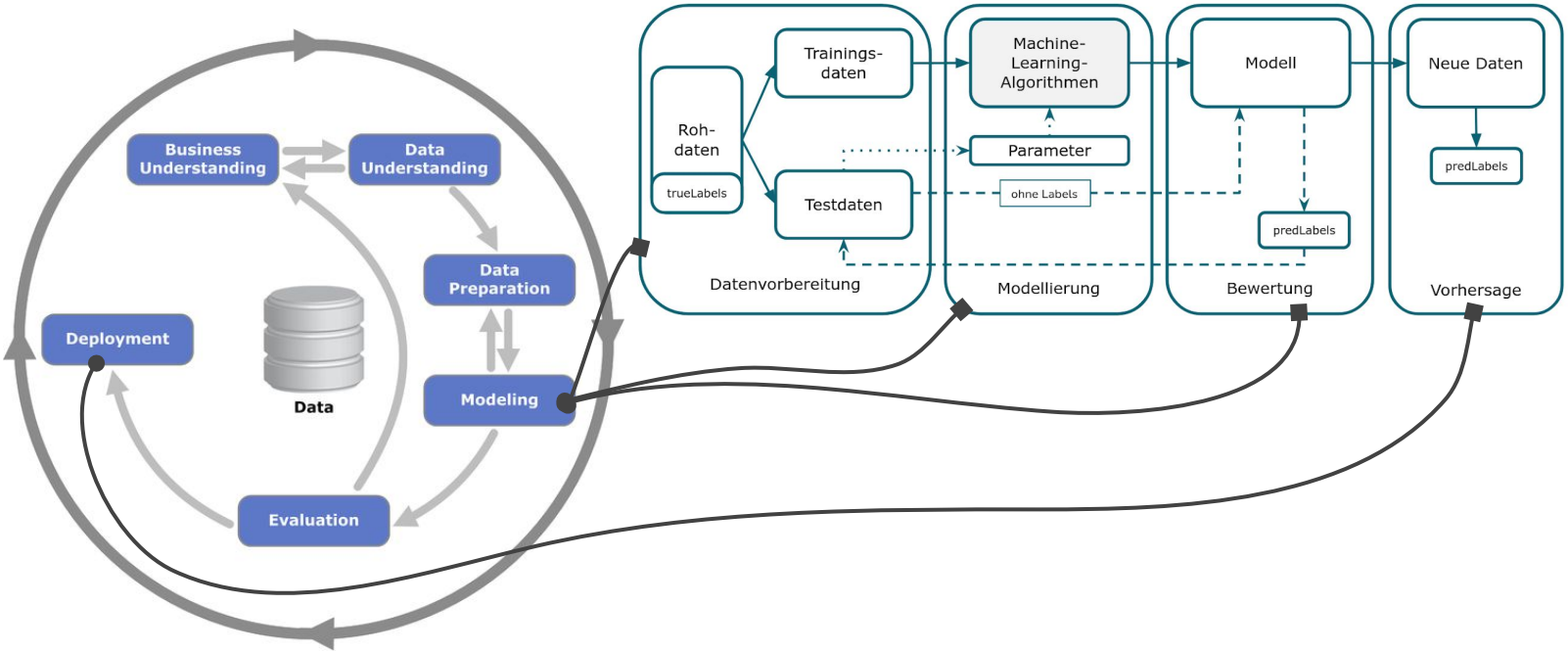
Quelle:
i.A.a. Raschka, Mirjalili (2018) Machine Learning with Python, mitp Verlag, 2. Auflage

ML-Prozess i.A.a. Raschka Überblick



Quelle:
Raschka, Mirjalili (2018) Machine Learning with Python, mitp Verlag, 2. Auflage

CRISP DM vs Raschka



Quelle:
https://en.wikipedia.org/wiki/File:CRISP-DM_Process_Diagram.png
i.A.a. Raschka, Mirjalili (2018) Machine Learning with Python, mitp Verlag, 2. Auflage

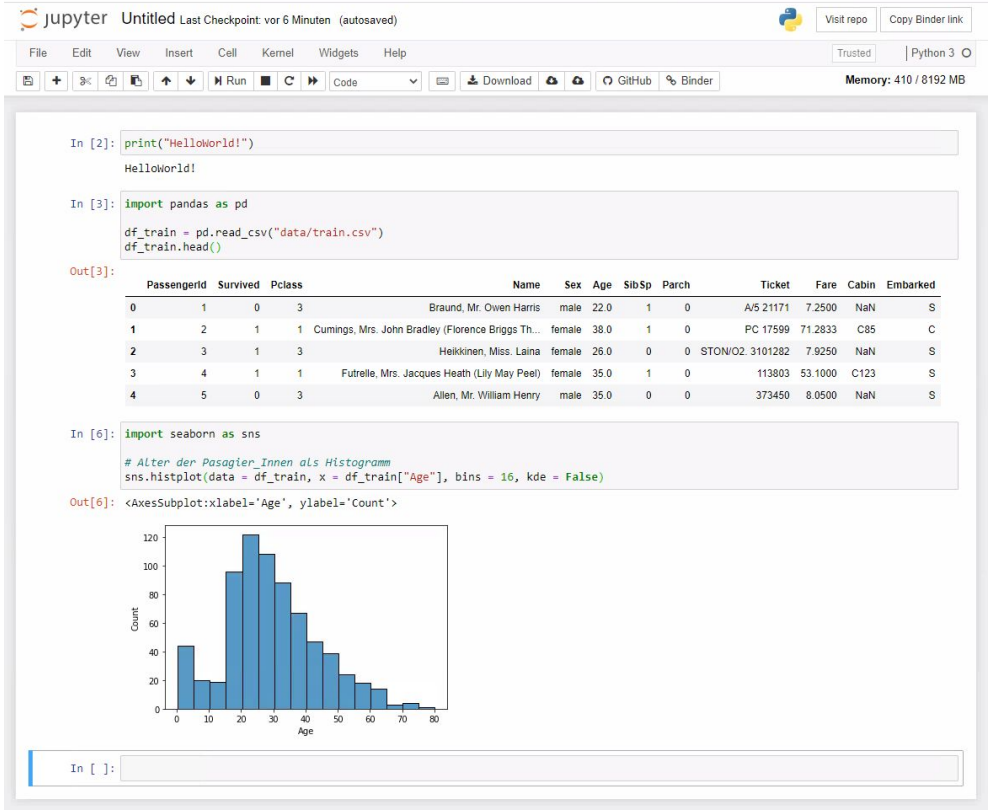
RealWeltBeispiele
Bundesdruckerei GmbH

- Werkzeugverschleiß
- Anomalieerkennung
- Prozessfähigkeit
- Bestelleingänge

Part 2

Jupyter Notebook 101

- Was ist so ein Notebook
- Wie ist es aufgebaut
- Wie verwende ich das



Titanic Dataset

Zwei Datensätze (train.csv, test.csv)
insgesamt 1309 Personen im DS (training 892, test 417)

Features:

- Passagier_In-Nummer
- Survived
- Pclass
- Name
- Sex
- Age
- SivSp
- Parch
- Ticket
- Fare
- Cabin
- Embarked

```
PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S
4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,35,1,0,113803,53.1,C123,S
5,0,3,"Allen, Mr. William Henry",male,35,0,0,373450,8.05,,S
6,0,3,"Moran, Mr. James",male,,0,0,330877,8.4583,,Q
7,0,1,"McCarthy, Mr. Timothy J",male,54,0,0,17463,51.8625,E46,S
8,0,3,"Palsson, Master. Gosta Leonard",male,2,3,1,349909,21.075,,S
9,1,3,"Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)",female,27,0,2,347742,11.1333,,S
10,1,2,"Nasser, Mrs. Nicholas (Adele Achem)",female,14,1,0,237736,30.0708,,C
11,1,3,"Sandstrom, Miss. Marguerite Rut",female,4,1,1,PP 9549,16.7,G6,S
12,1,1,"Bonnell, Miss. Elizabeth",female,58,0,0,113783,26.55,C103,S
13,0,3,"Saundercock, Mr. William Henry",male,20,0,0,A/5. 2151,8.05,,S
14,0,3,"Andersson, Mr. Anders Johan",male,39,1,5,347082,31.275,,S
15,0,3,"Vestrom, Miss. Hulda Amanda Adolfina",female,14,0,0,350406,7.8542,,S
16,1,2,"Hewlett, Mrs. (Mary D Kingcome) ",female,55,0,0,248706,16,,S
17,0,3,"Rice, Master. Eugene",male,2,4,1,382652,29.125,,Q
18,1,2,"Williams, Mr. Charles Eugene",male,,0,0,244373,13,,S
19,0,3,"Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)",female,31,1,0,345763,18,,S
20,1,3,"Masselmani, Mrs. Fatima",female,,0,0,2649,7.225,,C
21,0,2,"Fynney, Mr. Joseph J",male,35,0,0,239865,26,,S
22,1,2,"Beesley, Mr. Lawrence",male,34,0,0,248698,13,D56,S
23,1,3,"McGowan, Miss. Anna ""Annie""",female,15,0,0,330923,8.0292,,Q
24,1,1,"Sloper, Mr. William Thompson",male,28,0,0,113788,35.5,A6,S
```

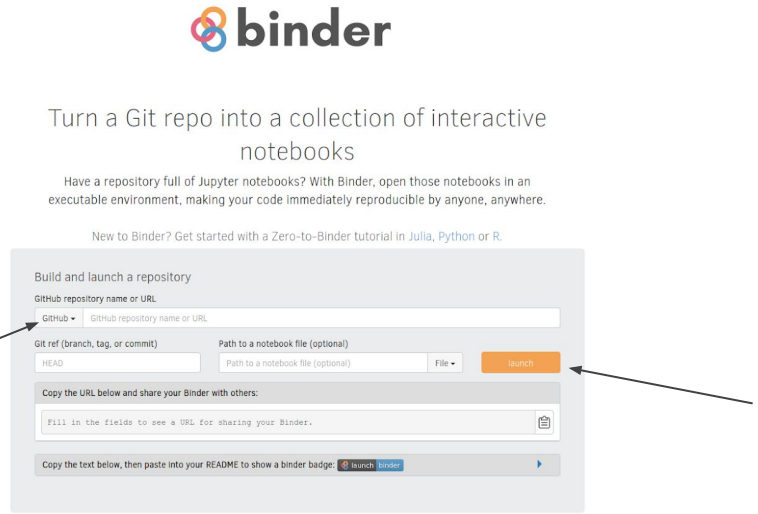
Titanic Dataset im Jupyter Notebook

Website aufrufen:
https://mybinder.org/

Repo Link einfügen:
https://github.com/dzwanzig/crisp_titanic

Launch

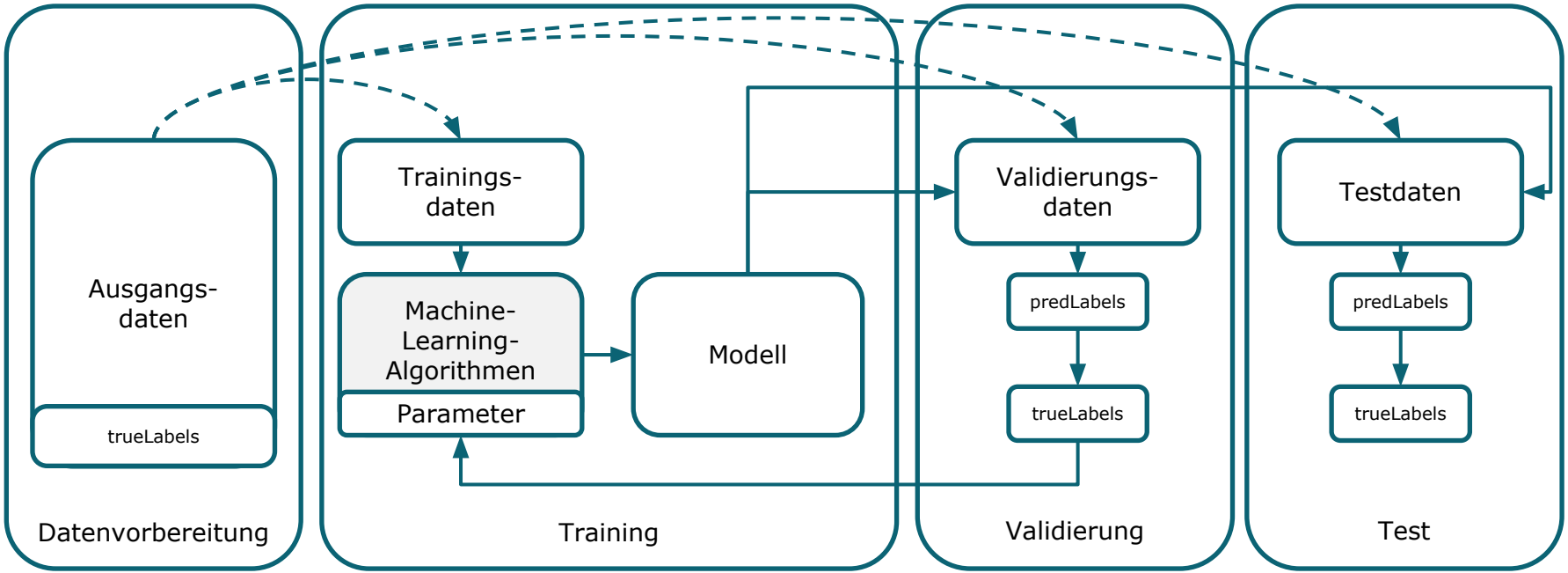
`Alternative:`
`https://mybinder.org/v2/gh/dzwanzig/_titanic/HEAD`



Vielen Dank

dorianzwanzig@gmail.com

ML-Prozess i.A.a. Raschka Überblick



Quelle:
i.A.a. Raschka, Mirjalili (2018) Machine Learning with Python, mitp Verlag, 2. Auflage