

DATA SCIENCE 2

VORLESUNG - NoCode

PROF. DR. CHRISTIAN BOCKERMANN

HOCHSCHULE BOCHUM

SOMMERSEMESTER 2023

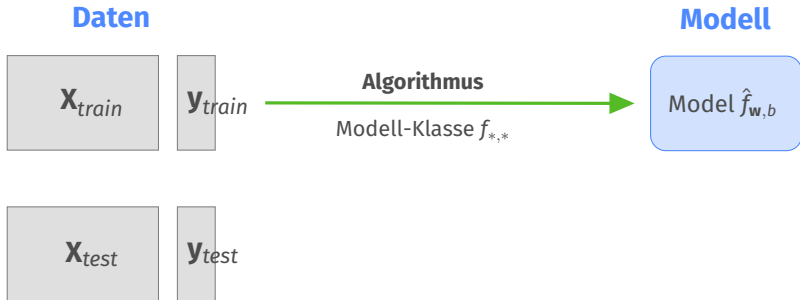
1 Datenanalyse mit Python

2 Weitere Software/Tools

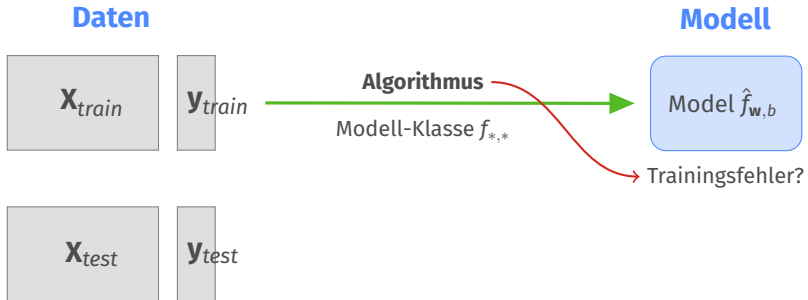
3 No-Code Ansätze

Datenanalyse mit Python

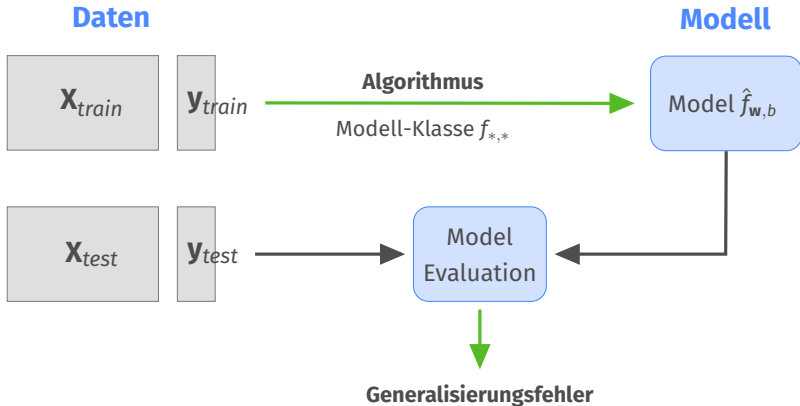
Vorgehen beim überwachten Lernen



Vorgehen beim überwachten Lernen



Vorgehen beim überwachten Lernen



```
import pandas as pd

# read data from csv
df = pd.read_csv('daten.csv')
features = ['a1', 'a2', 'a3']

# Merkmale auswaehlen
X = df[features]
y = df['label']

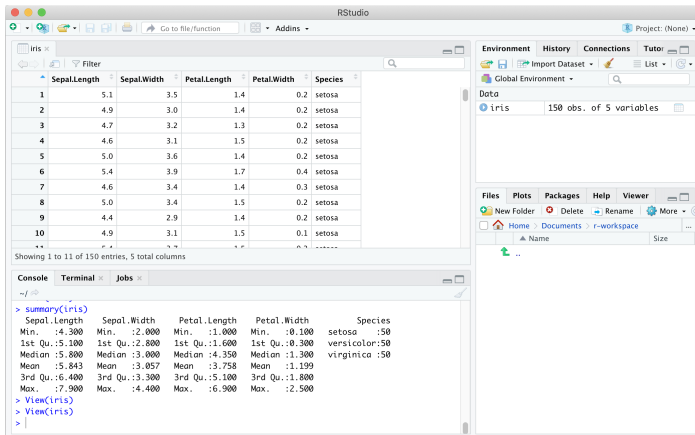
# Daten aufteilen
X_tr, X_ts, y_tr, y_ts = train_test_split(X, y)

# Modell trainieren
m = DecisionTreeClassifier()
m.fit(X_tr, y_tr)
```

Programmiersprachen

- Julia, <http://julialang.org>
- Python mit Pandas, SciKit Learn
<http://scikit-learn.org>
- R, <http://www.r-project.org>

Programmiersprache R für Statistik Aufgaben



The screenshot displays the RStudio interface with the following components:

- Environment:** Shows the 'iris' dataset with 150 observations and 5 variables.
- Files:** Shows the current workspace directory.
- Console:** Contains the following R code and output:

```
> summary(iris)
  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width  Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa   :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicol.:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica.:50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500

> View(iris)
> View(iris)
>
```

Abbildung: RStudio Umgebung für die Sprache R.

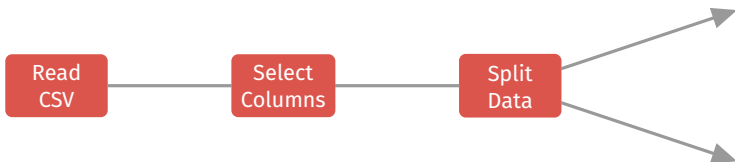
No-Code Ansätze

Trend: *No Code Tools*

- RapidMiner, <http://rapidminer.com>
- Knime, <http://www.knime.com>
- WEKA, MOA, <http://www.cs.waikato.ac.nz/ml/weka>
- Talend (Data Processing)

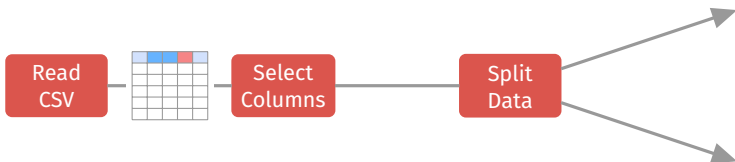
Werkzeuge um Prozesse mit graphischen Elementen zu entwerfen:

- Symbole für ausführbare Operationen
- Verbindungen zu Übergabe von Ergebnis-Objekten
- Einfache Start/Stopp Funktionen, Anhalten von Prozessen
- Möglichst ohne Programmierung auskommen



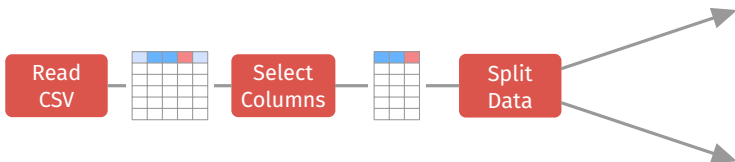
Werkzeuge um Prozesse mit graphischen Elementen zu entwerfen:

- Symbole für ausführbare Operationen
- Verbindungen zu Übergabe von Ergebnis-Objekten
- Einfache Start/Stopp Funktionen, Anhalten von Prozessen
- Möglichst ohne Programmierung auskommen



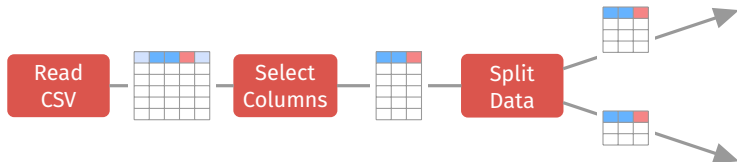
Werkzeuge um Prozesse mit graphischen Elementen zu entwerfen:

- Symbole für ausführbare Operationen
- Verbindungen zu Übergabe von Ergebnis-Objekten
- Einfache Start/Stopp Funktionen, Anhalten von Prozessen
- Möglichst ohne Programmierung auskommen



Werkzeuge um Prozesse mit graphischen Elementen zu entwerfen:

- Symbole für ausführbare Operationen
- Verbindungen zu Übergabe von Ergebnis-Objekten
- Einfache Start/Stopp Funktionen, Anhalten von Prozessen
- Möglichst ohne Programmierung auskommen



The screenshot displays the RapidMiner Studio Free 9.7.002 interface. The main workspace shows a process flow diagram with the following nodes: 'Daten Laden' (Data Load), 'Daten aufteilen' (Data Split), 'Modell trainieren' (Model Train), 'Modell anwenden' (Model Apply), and 'Performance'. The 'Modell trainieren' node is highlighted in orange. The interface includes a 'Repository' panel on the left with 'Import Data' and a tree view of resources. Below it is the 'Operators' panel showing categories like Modeling, Validation, and Utility. On the right, the 'Parameters' panel for 'Modell trainieren (Decision Tree)' is open, showing settings for 'criterion' (gain_ratio), 'maximal depth' (10), 'confidence' (0.1), 'minimal gain' (0.01), and 'minimal leaf size' (2). A 'Help' panel at the bottom right provides details about the 'Decision Tree' operator, including its category (Supervised Classification) and a synopsis stating it generates a decision tree.

Abbildung: Die graphische Schnittstelle von RapidMiner.

Prozesse werden als Graph mit vordefinierten Operator-Bausteinen gebaut

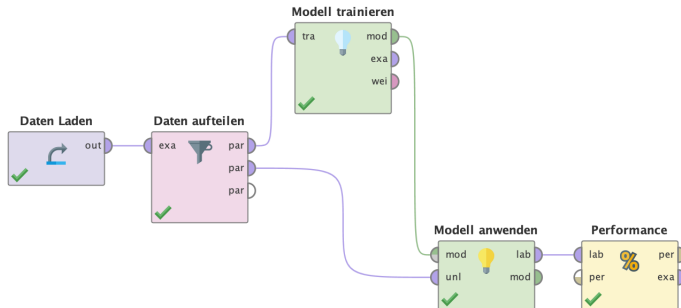


Abbildung: Ein Prozeß als Graph in RapidMiner.

RapidMiner wurde als OpenSource Tool am Lehrstuhl für künstliche Intelligenz der TU Dortmund entwickelt

- Prozess-Definition für ETL, Modellierung und Auswertung
- Einfaches Inspizieren / Exploration von Daten
- Enterprise Version für Unternehmen verfügbar
- Marktplatz mit Vielzahl von Erweiterungen
- *Wisdom of the crowds* Ansatz für schnellen Start

KNIME ist ebenfalls ein graphisches Tool für Prozess-Design

The screenshot displays the KNIME Analytics Platform interface. The main workspace shows a workflow titled "Visual Analysis of Sales Data". The workflow consists of the following nodes:

- Data Access:** "Provide file path" (File Reader) - Reads "sales_2008-2011.csv".
- Data Preprocessing:** "Filter columns and/or rows" (Column Filter) - Selects "country", "sales" and "amount" columns.
- Data Preprocessing:** "Filter columns and/or rows" (Row Filter) - Excludes rows where country is unknown.
- Data Visualization:** "Show sales by time and country" (Stacked Area Chart) - Displays "Sales by time".
- Data Visualization:** "Assign colors based on country" (Color Manager) - Assigns colors based on country.
- Data Visualization:** "Sales by country" (Pie/Donut Chart) - Displays "Sales by country".

The interface also includes a left sidebar with "KNIME Explorer" and "Node Repository", and a bottom panel with "Outline" and "KNIME Console". The console shows the following output:

```
=====
*** Welcome to KNIME Analytics Platform v4.2.2.v2820092508002 ***
*** Copyright by KNIME AG, Zurich, Switzerland ***
=====
Log file is located at: /Users/chris/.knime-workspace/.metadata/knime/knime.log
WARN Color Manager 3:2 Column "income" has no nominal values set
WARN Decision Tree Predictor 3:4 DataColumnSpec already contains a colo
WARN Decision Tree Predictor 3:4 DataColumnSpec already contains a colo
WARN Decision Tree Predictor 3:4 DataColumnSpec already contains a colo
WARN Decision Tree Predictor 3:4 DataColumnSpec already contains a colo
```

Abbildung: Die graphische Schnittstelle von KNIME.

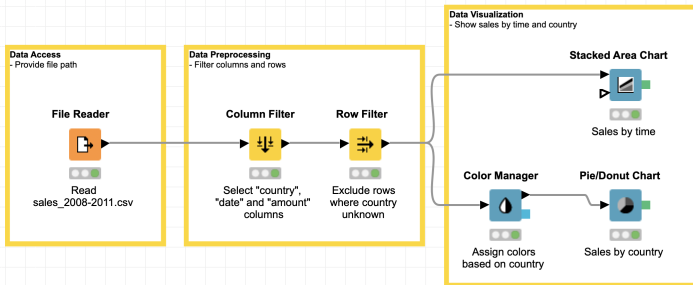


Abbildung: Ein Prozess zur Visualisierung mit KNIME.

Demo Rapidminer