DHBW CAS
Duale Hochschule Baden-Württemberg

Clustering

mit Scikit-Learn

Agenda

1. Definition Cluster-Analyse → Marius
2. Kontext Datensatz → Marius
3. Daten visualisieren & aufbereiten → Mario
4. Cluster-Analyse: kMeans → Marius
5. Cluster-Analyse: Hierarchisch → Mario
6. Kritische Reflexion → Marius
7. Fazit → Mario

Definition Cluster-Analyse

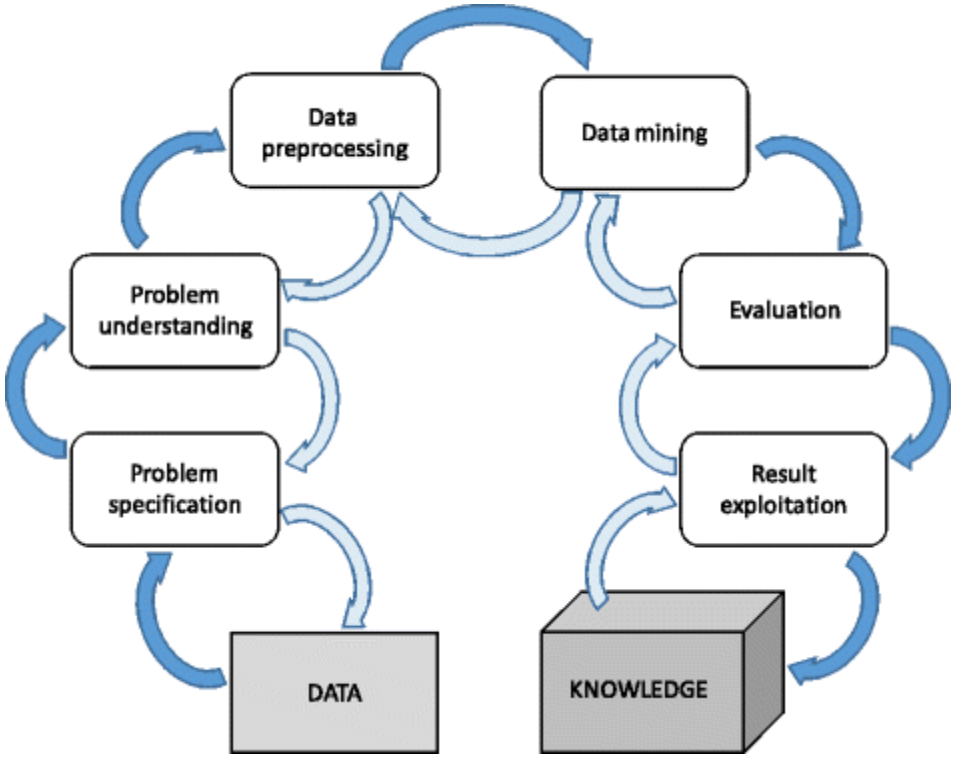
- Verfahren des maschinellen Lernens
- In einer Menge von **Daten** „ähnliche “ **Gruppierungen (Cluster)** erkennen
- Einsatz unterschiedlicher **Algorithmen** zur Bildung der Cluster

Kontext Datensatz

- **Raumklima**-Datensatz
- 15 Messungen mit Temperatur (°C) und Luftfeuchtigkeit (%)
- Unterschiedliche Kombinationen und damit Klima-Arten
- Gibt es ein **optimales Klima**?

Daten aufbereiten

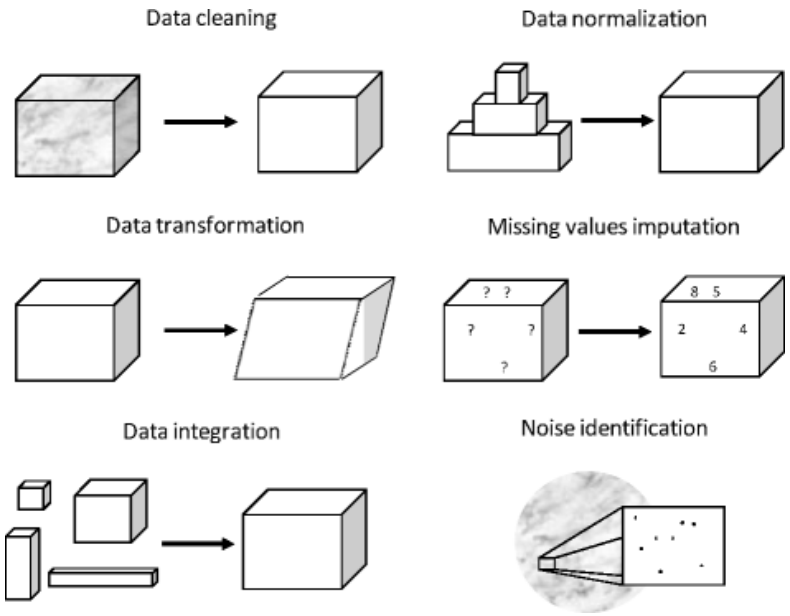
Entnommen aus: García, Salvador u. a. (2016)



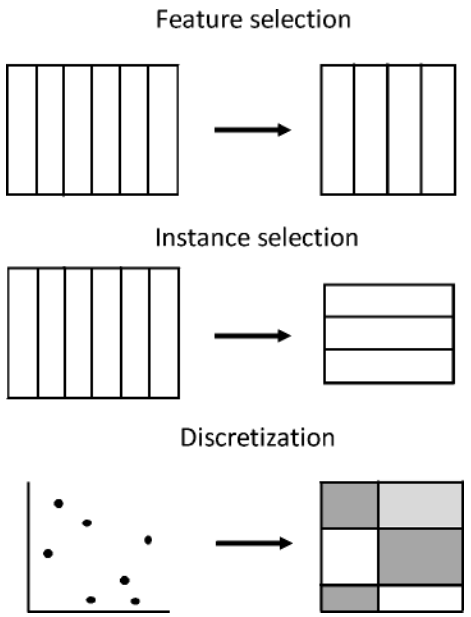
Knowledge Discovery in Databases - Prozess

Daten aufbereiten

Entnommen aus: García, Salvador u. a. (2016)

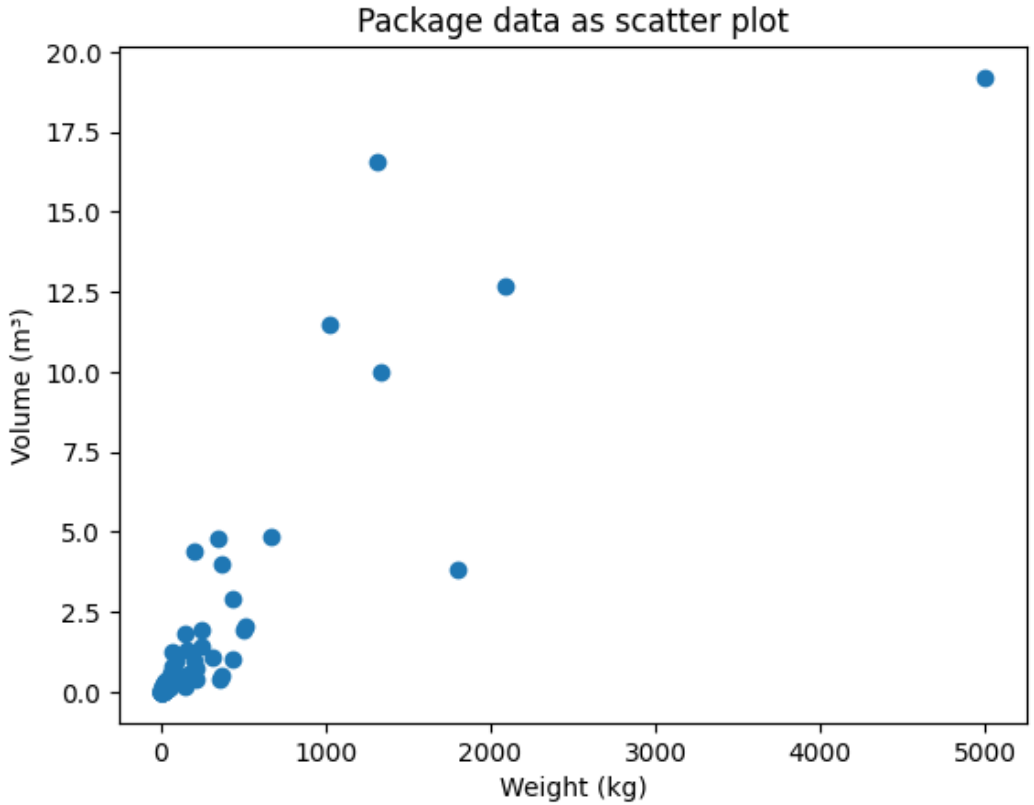


Data preprocessing tasks



Data reduction approaches

Daten visualisieren



Visualisiert mit Matplotlib

Daten aufbereiten

- In diesem Datensatz
  - **Data Cleaning**, bspw. 1.001,57 zu 1001.57
  - **Data Normalization**
  - **Data Integration** => Volumen ausrechnen
  - **Noise identification**
  - **Feature Selection**

Daten aufbereiten

#	Package No	Shipment No	Gross Weight (kg)	Width (cm)	Height (cm)	Length (cm)
0	1007530-2011-03239	1000088	23	35	30	35
1	1007530-2011-03241	1000310	150	60	55	80
2	1007530-2011-03242	1000346	0,5	14	15	19
3	1007530-2011-03243	1000456	1,5	20	20	29
4	1007530-2011-03244	1000796	1	10	10	10
5	1007530-2011-03245	1000957	75	82	81	120
6	1007530-2011-03246	1000957	41	80	34	120
7	1007530-2011-03247	1001184	1.340	220	112	406
8	1007530-2011-03249	1001408	0,5	20	20	29
9	1007530-2011-03250	1001563	5	45	35	45

Original-Datensatz (erste zehn Spalten)

#	Gross Weight (kg)	Width (cm)	Height (cm)	Length (cm)	Volume (cm^3)
0	23.0	35.0	30	35	36750.0
1	150.0	60.0	55	80	264000.0
2	0.5	14.0	15	19	3990.0
3	1.5	20.0	20	29	11600.0
4	1.0	10.0	10	10	1000.0
5	75.0	82.0	81	120	797040.0
6	41.0	80.0	34	120	326400.0
7	1340.0	220.0	112	406	10003840.0
8	0.5	20.0	20	29	11600.0
9	5.0	45.0	35	45	70875.0

Datensatz nach Aufbereitung (erste zehn Spalten)

Cluster-Analyse: kMeans

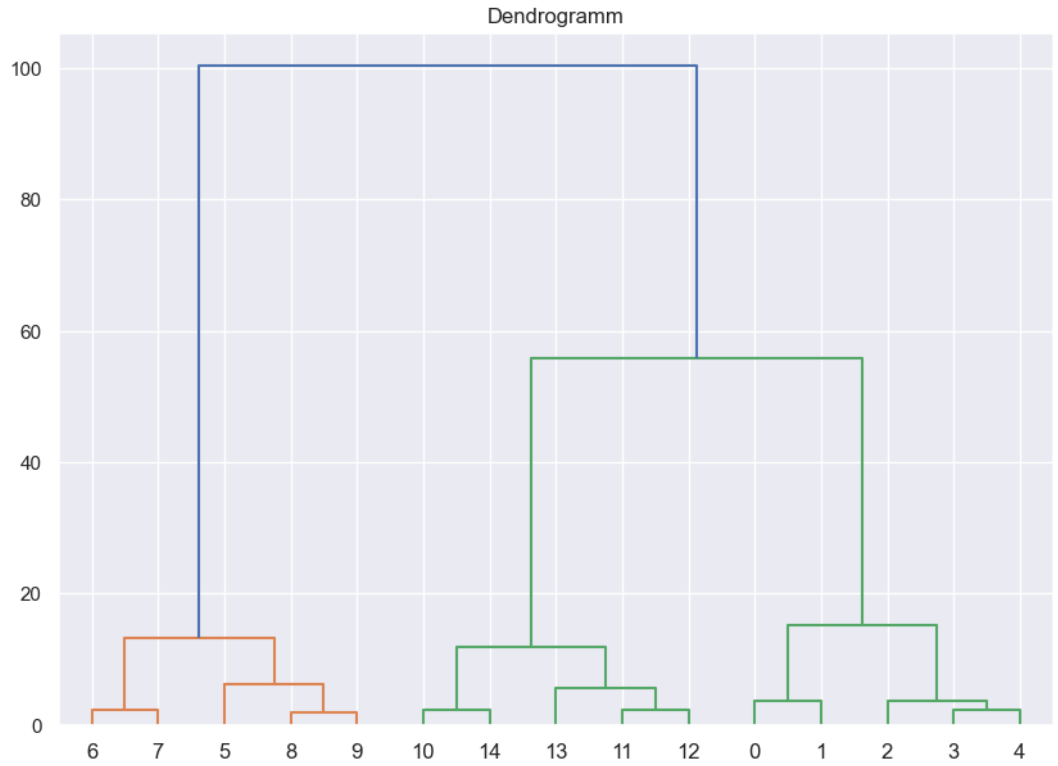
- Lorem

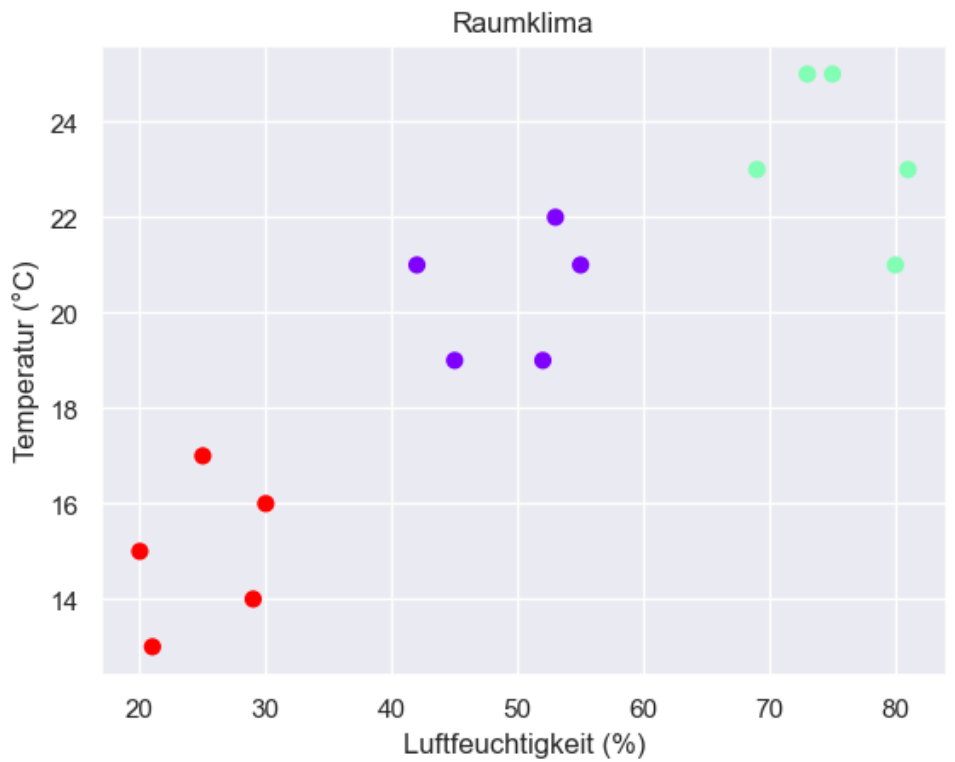
Cluster-Analyse: Hierarchisch

- **Agglomerative** Cluster-Analyse
- Darstellung in **Dendogramm**
- Abstandfunktion: **Euklidische** Distanz
- Fusionsvorschrift: **Ward** Methode

Cluster-Analyse: Hierarchisch

Visualisiert mit Matplotlib





Ausblick

- Anzahl Features (d): > 2
- Hyperparameter-Tuning
- Bias
- Over- & Underfitting
- Vergleich der unterschiedlichen Cluster-Scores

Quellen

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J.M. and Herrera, F., 2016. Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), pp.1-22.