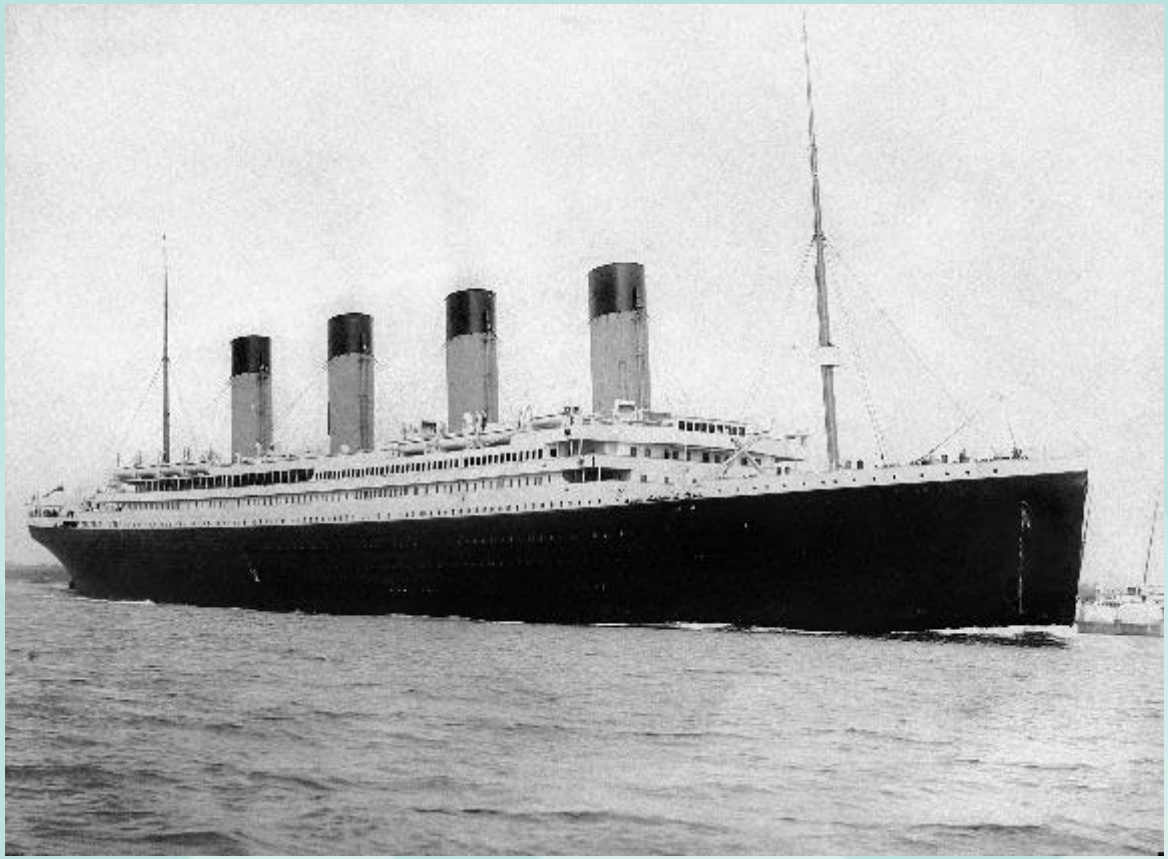


Big Data & Data Science

Beispieldatensätze

Prof. Dr. Klemens Waldhör

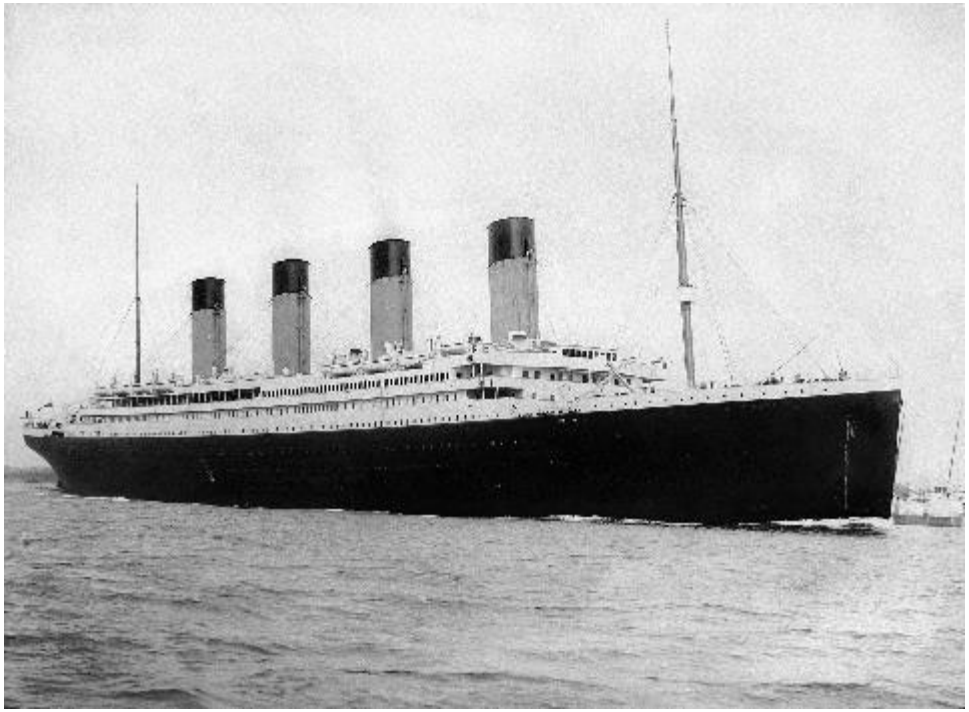
Überblick

Inhalt

- Beschreibung Titanic Datensatz
- Beschreibung Iris Datensatz

Ziel

- Einsatz in Übungsaufgaben



Titanic Datensatz

Titanic Datensatz

- Passagierschiff der britischen Reederei White Star Line
- Kollision Eisberg 14. April 1912 gegen 23:40 Uhr, gesunken 2 Stunden später
- 1514 der über 2200 an Bord befindlichen Personen sterben
- Gehört zu den größten und berühmtesten Katastrophen der Seefahrt

- Sehr beliebter Übungsdatensatz für Maschinelles Lernen
- Klassische Fragestellungen
  - **Wer überlebt den Untergang?**
  - **Welche Variablen beeinflussen die Überlebenswahrscheinlichkeit**

Titanic und R:
https://sebastiansauer.github.io/titanic/

Titanic Datensatz

- Datei: Titanic\titanic_exported.csv

```
"PassengerId";"Survived";"Pclass";"Name";"Sex";"Age";"SibSp";"Parch";"Ticket";"Fare";"Cabin";"Embarked"
1;0;3;"Braund, Mr. Owen Harris";"male";22.0;1;0;"A/5 21171";7.25;;"S"
2;1;1;"Cumings, Mrs. John Bradley (Florence Briggs Thayer)";"female";38.0;1;0;"PC 17599";71.2833;"C85";"C"
3;1;3;"Heikkinen, Miss. Laina";"female";26.0;0;0;"STON/O2. 3101282";7.925;;"S"
4;1;1;"Futrelle, Mrs. Jacques Heath (Lily May Peel)";"female";35.0;1;0;"113803";53.1;"C123";"S"
5;0;3;"Allen, Mr. William Henry";"male";35.0;0;0;"373450";8.05;;"S"
```

- **PassengerId – interne ID (1…891)**

SURVIVED	Überlebensvariable mit 0=Nein und 1=Ja
PCLASS	Klasse auf dem Schiff 1= Erste, 2=Zweite, 3= Dritte
NAME	Nachname, Title, Vorname(n)
SEX	Geschlecht „male“=männlich, „female“=weiblich
AGE	Alter in Jahren
SIBSP	Anzahl der Geschwister bzw. Partner, die mit an Bord waren
PARCH	Anzahl der Eltern bzw. Kinder, die mit an Bord waren
TICKET	ID Nummer des Tickets
FARE	Preis des Tickets
CABIN	Kabinennummer
EMBARKED	Einstiegshafen C= Cherbourg, Q= Queenstown, S=Southampton

Titanic Datensatz

- Python Beispiel Datensatz laden in Jupiter
- scikit-learn Package

```
3]: import numpy as np
from sklearn.datasets import fetch_openml
np.random.seed(42)
X, y = fetch_openml("titanic", version=1, as_frame=True, return_X_y=True)
# X.drop(['boat', 'body', 'home.dest'], axis=1, inplace=True)
# X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y, test_size=0.2)
```

```
4]: X
```

4]:

	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	1.0	Allen, Miss. Elisabeth Walton	female	29.0000	0.0	0.0	24160	211.3375	B5	S	2	NaN	St Louis, MO
1	1.0	Allison, Master. Hudson Trevor	male	0.9167	1.0	2.0	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
2	1.0	Allison, Miss. Helen Loraine	female	2.0000	1.0	2.0	113781	151.5500	C22 C26	S	None	NaN	Montreal, PQ / Chesterville, ON
3	1.0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1.0	2.0	113781	151.5500	C22 C26	S	None	135.0	Montreal, PQ / Chesterville, ON
4	1.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1.0	2.0	113781	151.5500	C22 C26	S	None	NaN	Montreal, PQ / Chesterville, ON
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1304	3.0	Zabour, Miss. Hileni	female	14.5000	1.0	0.0	2665	14.4542	None	C	None	328.0	None
1305	3.0	Zabour, Miss. Thamine	female	NaN	1.0	0.0	2665	14.4542	None	C	None	NaN	None
1306	3.0	Zakarian, Mr. Mapriededer	male	26.5000	0.0	0.0	2656	7.2250	None	C	None	304.0	None
1307	3.0	Zakarian, Mr. Ortin	male	27.0000	0.0	0.0	2670	7.2250	None	C	None	NaN	None
1308	3.0	Zimmerman, Mr. Leo	male	29.0000	0.0	0.0	315082	7.8750	None	S	None	NaN	None

1309 rows × 13 columns

```
5]: y
```

```
5]: 0       1
1       1
2       0
3       0
4       0
       ..
1304    0
1305    0
1306    0
1307    0
1308    0
Name: survived, Length: 1309, dtype: category
Categories (2, object): ['0', '1']
```

https://jaketae.github.io/study/sklearn-pipeline/

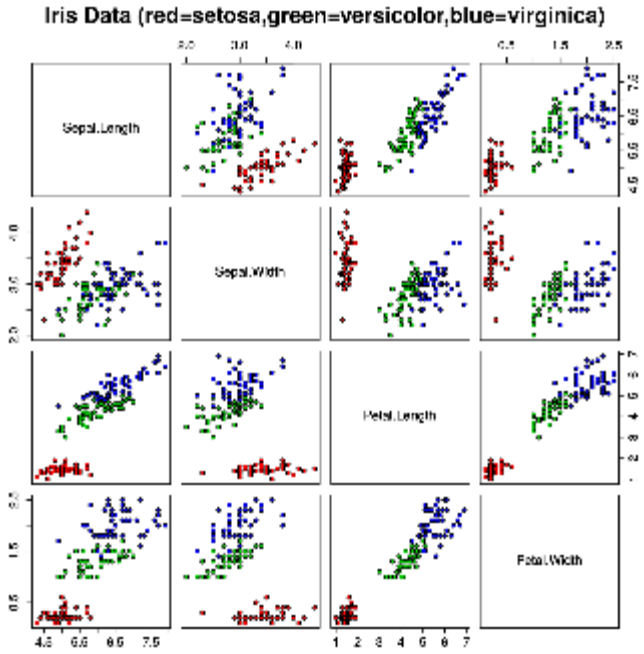
Iris Datensatz

Iris Datensatz

- Fisher‘s Iris Datensatz
- 150 Beobachtungen von 4 Attributen von Schwertlilien
- https://www.wanderinformatiker.at/unipages/general/iris.html
- „Beim Iris Datensatz… handelt es sich um einen Datensatz mit 150 Beobachtungen von 4 Attributen von Schwertlilien. Gemessen wurden dabei jeweils die Breite und die Länge des Kelchblatts (Sepalum) sowie des Kronblatts (Petalum) in Zentimeter. Des weiteren ist für jeden Datensatz die Art der Schwertlilie (Iris setosa, Iris virginica oder Iris versicolor) angegeben.“

- Sehr beliebter Übungsdatensatz für Maschinelles Lernen

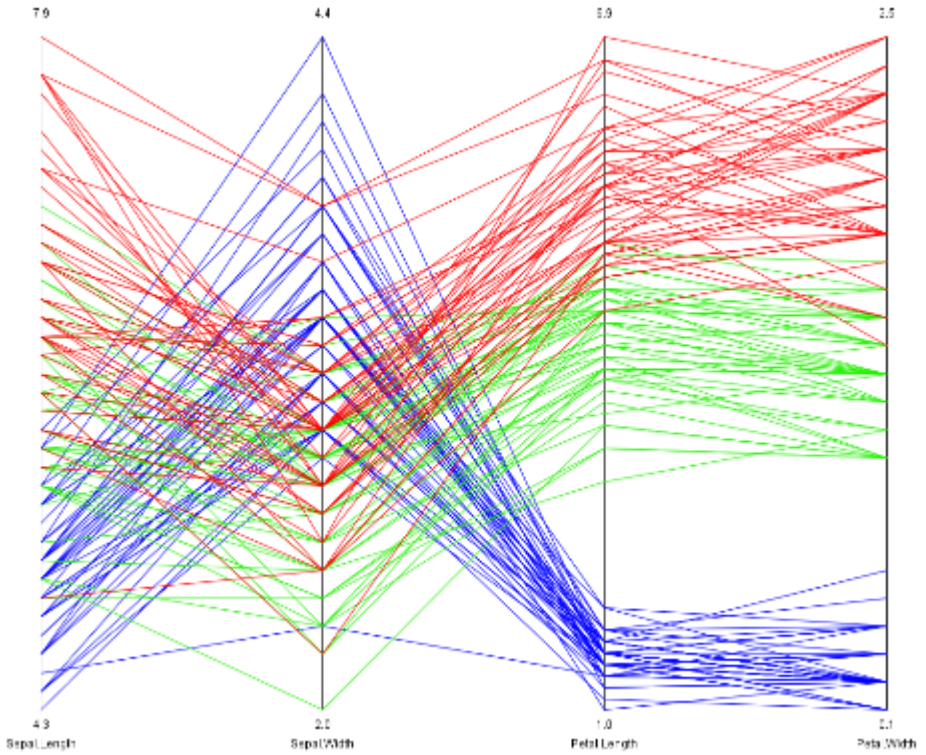
- https://archive.ics.uci.edu/ml/datasets/Iris
- https://www.kaggle.com/uciml/iris
- https://en.wikipedia.org/wiki/Iris_flower_data_set
- http://blueowlpress.com/wp-content/uploads/IrisTutorialPart1.html



Iris Datensatz

- Datei: Iris\iris.data
- 150 Datensätz
- Struktur und Variablen

```
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
        -- Iris Setosa
        -- Iris Versicolour
        -- Iris Virginica
```



- Beispiel

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
…
```

Iris Datensatz

- Python Beispiel Datensatz laden in Jupiter
- scikit-learn Package

```
[3]: from sklearn.datasets import load_iris
     iris = load_iris()
     iris
```

```
[3]: {'data': array([[5.1, 3.5, 1.4, 0.2],
                    [4.9, 3. , 1.4, 0.2],
                    [4.7, 3.2, 1.3, 0.2],
                    [4.6, 3.1, 1.5, 0.2],
                    [5. , 3.6, 1.4, 0.2],
                    [5.4, 3.9, 1.7, 0.4],
                    [4.6, 3.4, 1.4, 0.3],
                    [5. , 3.4, 1.5, 0.2],
                    [4.4, 2.9, 1.4, 0.2],
                    [4.9, 3.1, 1.5, 0.1],
                    [5.4, 3.7, 1.5, 0.2],
                    [4.8, 3.4, 1.6, 0.2]
```

Kriminalität USA

Kriminalität USA Datensatz

- State Crime CSV File
- Daten zu verschiedenen Delikten in den USA
- https://corgis-edu.github.io/corgis/csv/state_crime/