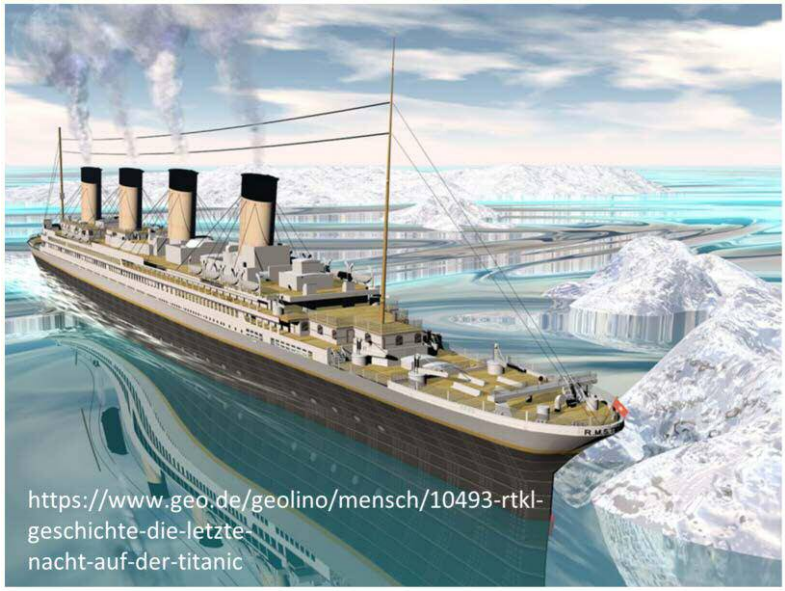
Metadaten Datensatz 1: Titanic - Überleben einer Katastrophe



Daten	**Titanic.zip**
Inhalte	Passagierliste der Titanic (train_with_survival_info.csv, test_no_survival_info.csv)
Informationen	Der Train-Datensatz beinhaltet die Informationen bzgl. Überlebensstatus. Mit diesem Datensatz können Sie das Modell entwickeln. Der Test-Datensatz soll genutzt werden, um die Performance des Modells zu testen. Hier gibt es keine Informationen zum Überlebensstatus der Passagiere und das Modell soll diesen Vorhersagen.
Fragestellung	Welche Passagiere im Testdatensatz können die Katastrophe überleben und welche nicht? Welche Passagiermerkmale haben den größten Einfluss auf die Überlebenschancen?
Vokabular	Survival - Überlebensstatus 0 = Nein, 1 =Ja pclass Ticket Klasses 1 = Erste Klasse (Upper) , 2 = Zweite Klasse, 3 = Dritte Klasse (zur Einschätzung des sozio-ökonomischen Status der Passagiere) Name sex Geschlecht male /female Age Alter (Jahre) sibsp Anzahl der Geschwister / Ehepartner an Bord der Titanic parch Anzahl der Eltern / Kinder an Bord der Titanic ticket Ticket-Nummer fare Ticketpreis cabin Kabinen-Nummer embarked Zustiegshafen C = Cherbourg, Q = Queenstown, S = Southampton
Quelle	https://www.kaggle.com/competitions/titanic

Metadaten Datensatz 2: Diabetes – Diagnostische Vorhersage einer Erkrankung

https://www.hescuro.de/tests/diabetes-test/

Daten	**diabetes.csv**
Inhalte	Patientenliste vom US National Institute of Diabetes and Digestive and Kidney Diseases
Informationen	Dieser Datensatz enthält die medizinischen Messdaten von Patientinnen, die von nordamerikanischen Indigenen abstammen (Pima-Indianer). Diese Gruppe zeigt ein besonders erhöhtes Risiko, an Diabetes zu erkranken.
Fragestellung	Ziel des Datensatzes ist es, auf der Grundlage bestimmter diagnostischer Merkmale vorherzusagen, ob eine Patientin am Beginn einer Diabeteserkrankung steht oder nicht. Welche Patientinnen-Merkmale haben den größten Einfluss auf das Erkrankungsrisiko?
Vokabular	Medizinische Variable (Prädiktoren) im Datensatz sind: • Pregnancies = Anzahl der Schwangerschaften • Glucose = Glukosekonzentration nach Glukosetoleranztest • BloodPressure = Blutdruck (mg Hg) • SkinThickness = Dicke der Trizepshautfalte (mm) • Insulin = Insulin (µU/ml) • BMI = Body Mass Index (in kg/Körperhöhe in m)^2 • DiabetesPedigreeFunction = Diabetes Stammbaumfunktion (bewertet die Wahrscheinlichkeit von Diabetes aufgrund der Familienanamnese) • Age = Alter (Jahre) • Outcome = nicht erkrankt (=0), erkrankt (=1)
Quelle	• Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press. • https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Metadaten Datensatz 3: Immobilienpreise – welche Eigenschaften des Umfeldes sind entscheidend?

Daten	**Hauspreise_Boston.csv**
Inhalte	In diesem Datensatz sind Informationen über die Preise von über 500 Häusern in Boston (USA) enthalten. Zusätzlich finden sich im Datensatz charakteristische Merkmale zum Umfeld der Häuser, wie z.B. Kriminalitätsrate, Luftqualität oder Bevölkerung.
Informationen	Der Datensatz dient der Analyse von Einflussfaktoren auf amerikanische Immobilienpreise.
Fragestellung	Mit welcher Genauigkeit lassen sich aus charakteristischen Umfeldmerkmalen die Hauspreise vorhersagen? Welche Merkmale sind die stärksten Einflussfaktoren?
Vokabular	• CRIME: pro Kopf Kriminalitätsrate • ZN: Anteil der Wohnbauflächen für Grundstücke mit einer Fläche von mehr als 25.000 m² • INDUS: Anteil der Nicht-Einzelhandelsflächen • RIVER: Dummy Variable für Fluss (1 – grenzt an Fluss, 0 - sonst) • AIRQ: Luftqualitätsindikator NOx Konzentration in parts per 10 million • RM: durchschnittliche Anzahl der Zimmer pro Wohnung • AGE: Anteil der vor 1940 gebauten selbstgenutzten Wohneinheiten • DIS: gewichtete Entfernungen zu fünf Beschäftigungszentren • RAD: Index der Erreichbarkeit von Autobahnen und Schnellstraßen • TAX: Vollwertiger Grundsteuersatz pro 10.000 $ • PTRATIO: Index für Schüler-Lehrer Verhältnis • CPRATIO: Index zur ethnischen Zusammensetzung der Bevölkerung • LSTAT: Prozentanteil Bevölkerung mit niedrigem Lebensstatus • PRICE: gemittelter der Eigenheime in US$
Quelle	• The Boston house-price data of Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978. • https://www.kaggle.com/code/alisonmachadoluiz/boston-house-prices-multiple-regressions/input