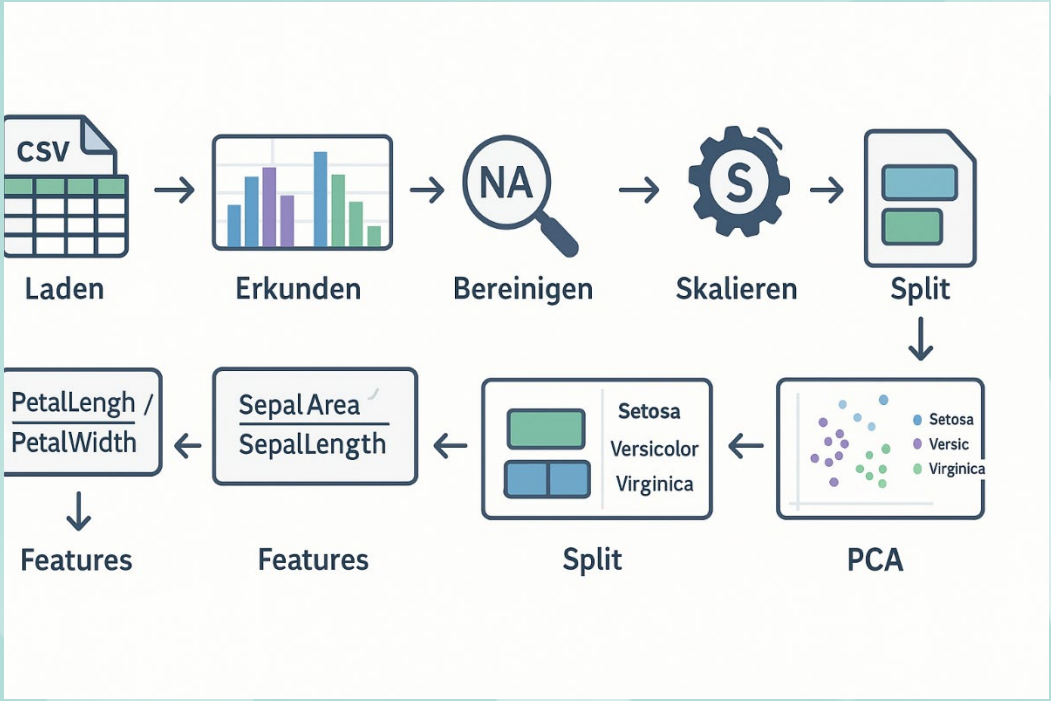
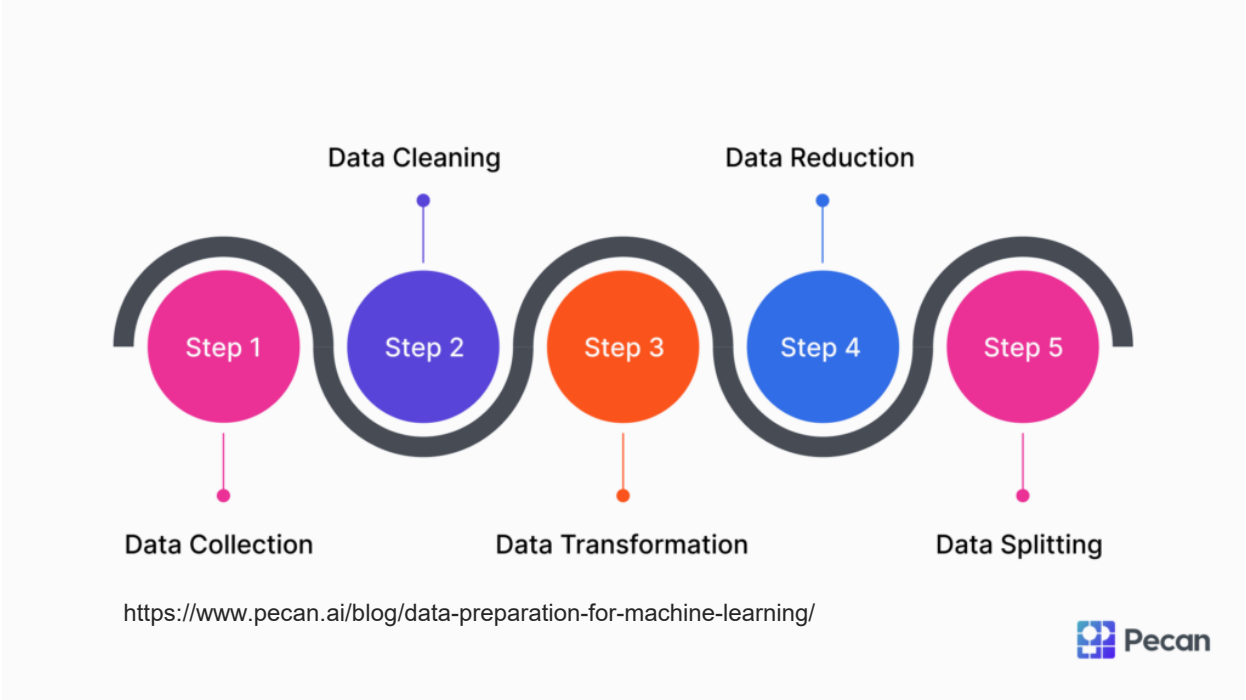


Big Data and Data Science
Data Preparation

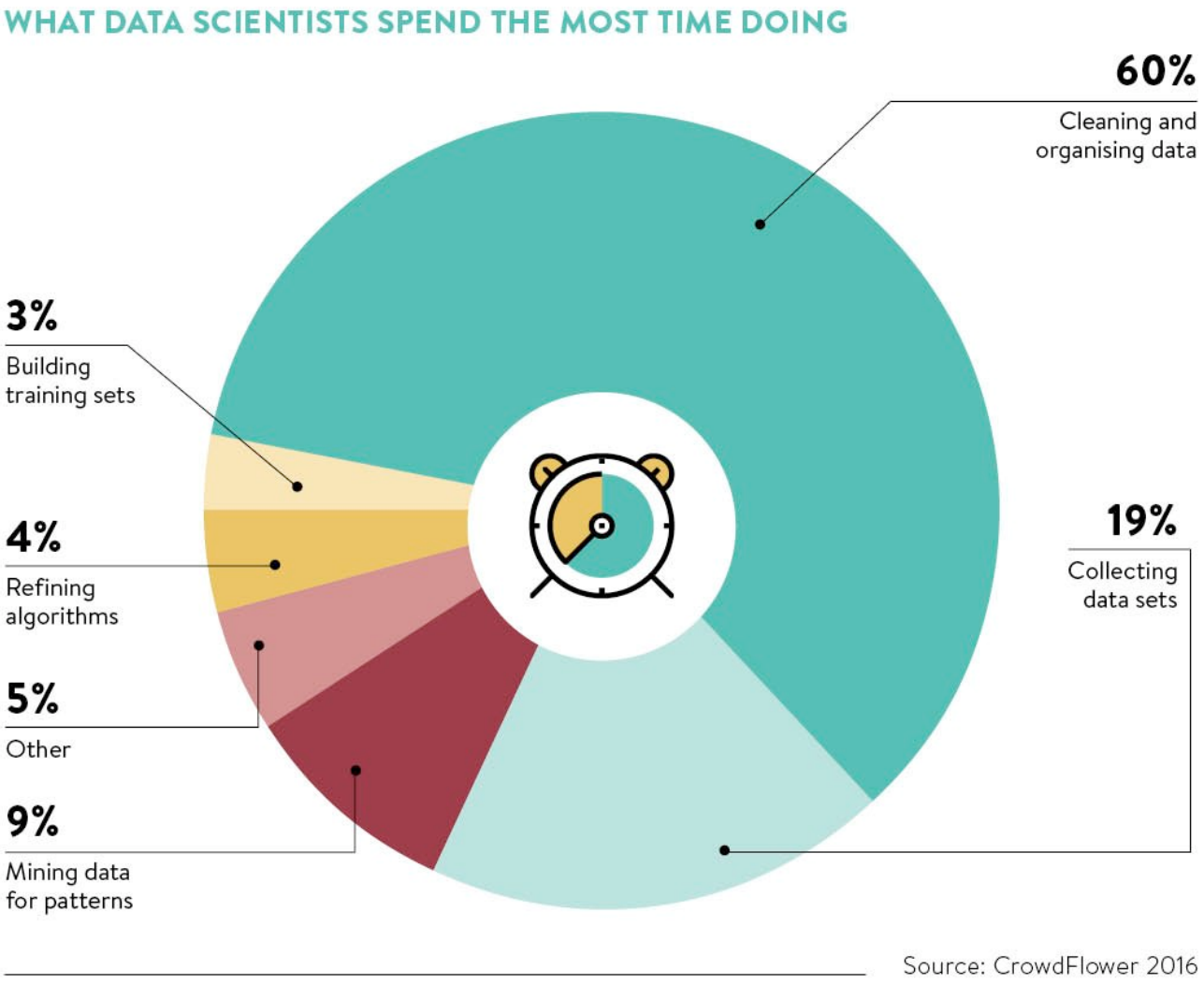
WS 2025/26

Prof. Dr. Klemens Waldhör.

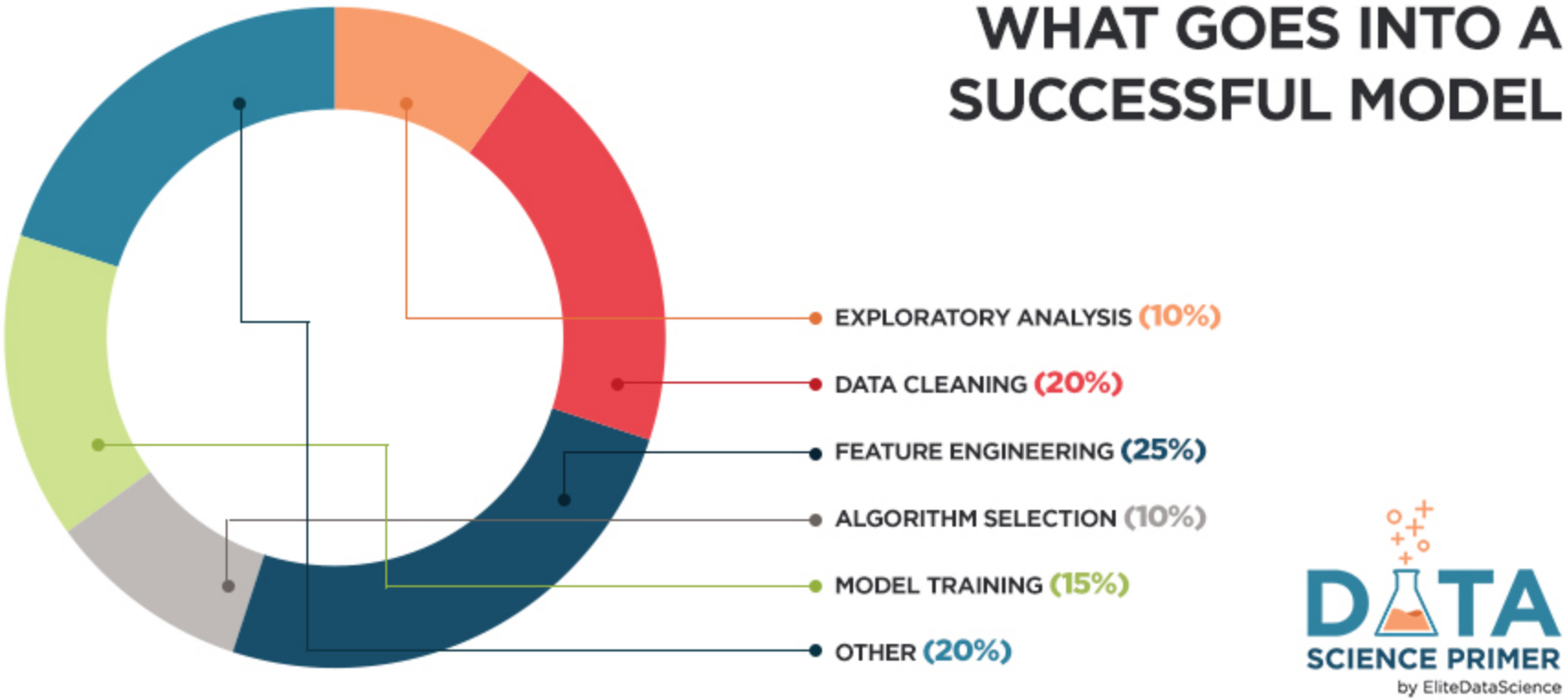


1. Einführung

Einführung


WHAT DATA SCIENTISTS SPEND THE MOST TIME DOING
60%
Cleaning and organising data
3%
Building training sets
4%
Refining algorithms
5%
Other
9%
Mining data for patterns
19%
Collecting data sets
Source: CrowdFlower 2016

Einführung



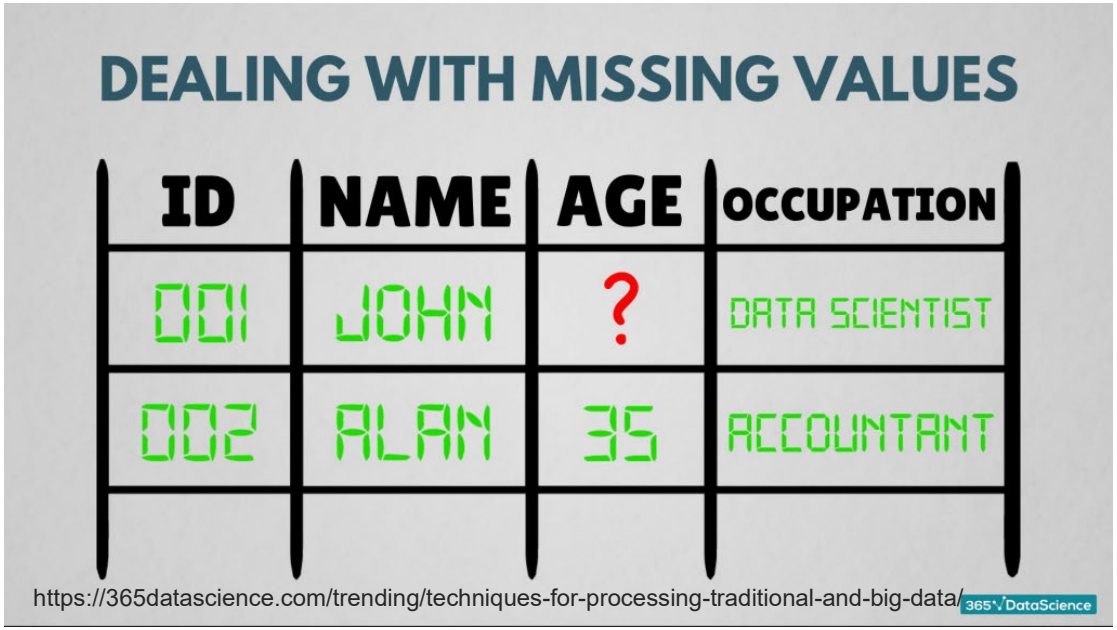
https://elitedatascience.com/feature-engineering

Einführung

- Data Preparation bezeichnet den Prozess, bei dem Rohdaten in eine geeignete Form gebracht werden, um sie für Machine-Learning-Modelle nutzbar zu machen.
- Es ist ein entscheidender Schritt im gesamten ML-Workflow, da die Qualität der vorbereiteten Daten maßgeblich die Leistung des Modells beeinflusst.
- Die Daten sollen strukturiert, konsistent und modellkompatibel sein – damit das Machine-Learning-Modell Muster zuverlässig erkennen und generalisieren kann.

Wichtige Schritte der Data Preparation

- Datenbereinigung
  Entfernen von Duplikaten, fehlerhaften oder fehlenden Werten.
- Datenintegration
  Zusammenführen verschiedener Datenquellen zu einem konsistenten Datensatz.
- Feature Engineering
  Erstellen neuer Merkmale oder Transformation bestehender, um die Modellleistung zu verbessern.
- Normalisierung/Skalierung
  Anpassung von Wertebereichen, z. B. durch Min-Max-Skalierung oder Standardisierung.
- Kodierung kategorialer Variablen
  Umwandlung von Textdaten in numerische Formate (z. B. One-Hot-Encoding).
- Datenaufteilung
  Trennung in Trainings-, Validierungs- und Testdaten.



2. Behandlung „Missing Values“

Handling „Missing Values“

- Eine der häufigsten Unregelmäßigkeiten in Datensätzen sind fehlende Merkmalsausprägungen **(„Missing Values“)**
- In RapidMiner gekennzeichnet durch „?“
- Sollte die Daten Exploration aufzeigen, dass der Datensatz „Missing Values“ enthält bietet sich folgendes Vorgehen an:

1. Ursachenforschung

- Ist das Fehlen der Werte zufallsbedingt?
- Bestand ein Problem in der Aufzeichnung, Übertragung oder Speicherung der Daten?
- Lassen sich die fehlenden Werte noch aus einer anderen Quelle wiederherstellen bzw. nachträglich erfassen?

2. Behandlung von „Missing Values“

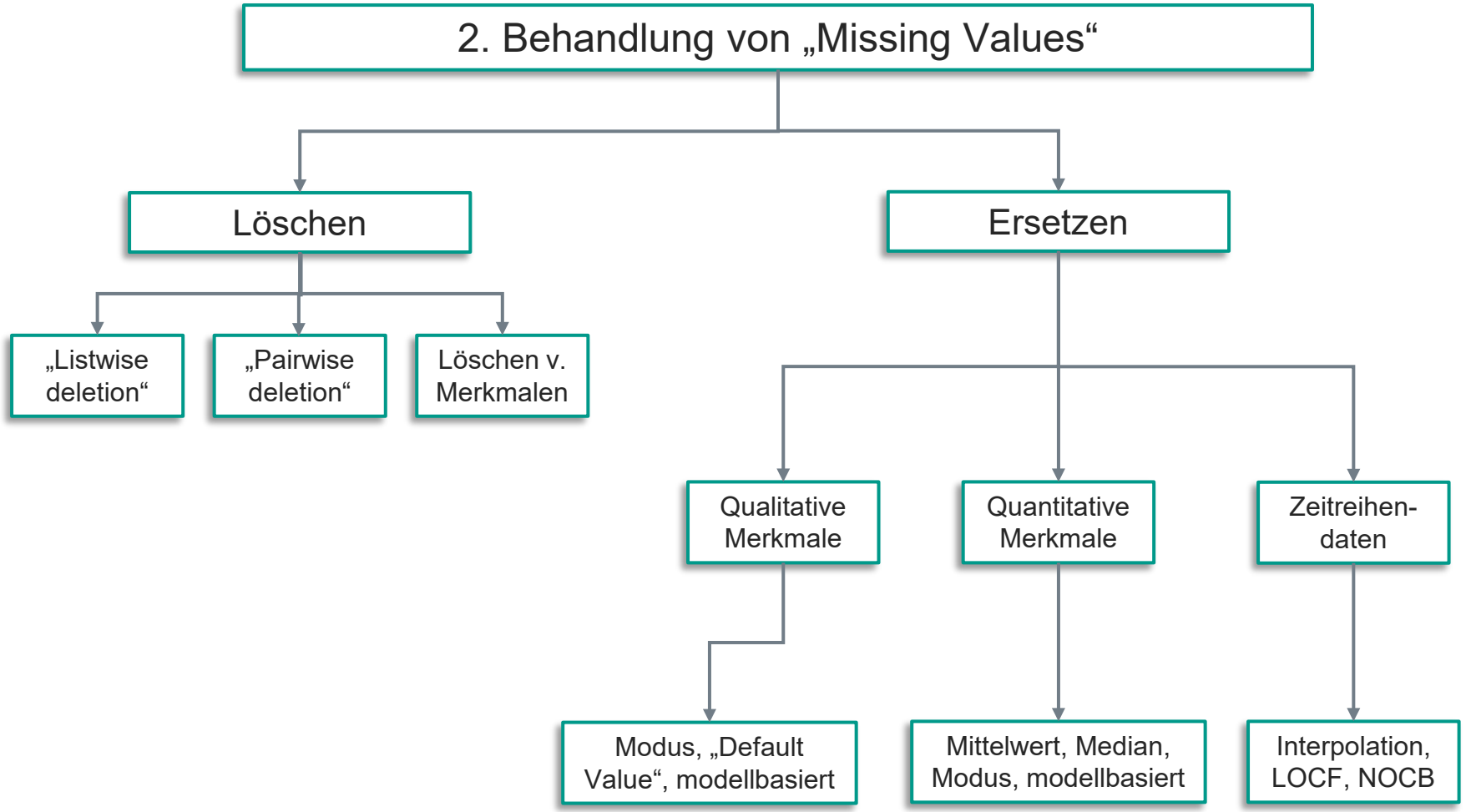
- Löschen einzelner Werte, Observationen oder Merkmale
- Ersetzen fehlender Werte

Handling „Missing Values“

1. Ursachenforschung

- Arten von „Missing Values“:
  - **Missing completely at random (MCAR)**
    - Wahrscheinlichkeit eines fehlenden Wertes ist für alle Merkmalsausprägungen gleich z.B. Die Altersangabe fehlt bei einigen Passagieren, jedoch unabhängig von den Ausprägungen anderer Merkmale
    - Der Idealfall, die fehlenden Werte treten rein zufällig auf.
  - **Missing at random (MAR)**
    - Wahrscheinlichkeit eines fehlenden Wertes ist höher, wenn ein anderes Merkmal eine bestimmte Ausprägung besitzt. z.B. Bei Passagieren der 3. Klasse fehlt die Altersangabe häufiger als bei Passagieren der 1. Klasse
    - Überprüfbar durch Visualisierung der „Missing Value“-Verteilung über die Merkmalsausprägungen
  - **Missing not at random (MNAR)**
    - Wahrscheinlichkeit eines fehlenden Wertes ist für einige Merkmalsausprägungen höher und bedingt durch das Merkmal selbst z.B.: Passagiere, die einen hohen Preis für ihr Ticket bezahlt haben unterschlagen den Preis bewusst
    - Nur feststellbar durch intensive Auseinandersetzung mit der Datenerfassungsmethodik
    - Löschen/Ersetzen der Werte ist nicht möglich ohne dabei das Modell zu beeinflussen.

Handling „Missing Values“



2. Behandlung von „Missing Values“

Löschen von „Missing Values“:

- **„Listwise deletion“**
  - Löschen aller Observationen, für die in mindestens einem Merkmal fehlende Werte vorliegen
  - Anzahl der Observationen pro Merkmal bleibt für alle Merkmale identisch
- **„Pairwise deletion“**
  - Observationen werden bei der Betrachtung von Merkmalen, für die sie keine Ausprägung besitzen, ignoriert.
  - Führt zu einer unterschiedlichen Anzahl an Observationen pro Merkmal
- **Löschen ganzer Merkmale**
  - Enthält ein Merkmal zu viele fehlende Ausprägungen (>50%), ist es sinnvoll zu prüfen, ob das gesamte Merkmal aus der Analyse entfernt werden kann.

Handling „Missing Values“

2. Behandlung von „Missing Values“

Ersetzen Qualitativer und Quantitativer Merkmalsausprägungen:

- **Schätzen fehlender Werte auf Grundlage aller Ausprägungen eines Merkmals**
  - Ersetzen der „Missing Values“ durch Modus, Median oder Mittelwert über alle Ausprägungen
  - Sinnvoll bei „MCAR Values“
  - Nachteil: Verschiebt ggf. Varianz des Merkmals

- **Schätzen fehlender Werte auf Grundlage ähnlicher Observationen**
  - Ersetzen der „Missing Values“ durch Modus, Median oder Mittelwert der Ausprägungen ähnlicher Observationen
  - z.B. Ersetzen fehlender Altersangaben durch den Mittelwert getrennt nach Geschlechtern
  - Sinnvoll bei „MAR Values“

Handling „Missing Values“

2. Behandlung von „Missing Values“

Ersetzen Qualitativer und Quantitativer Merkmalsausprägungen:

- **Modellbasiert**
  - Verwenden eines Klassifikations-/Regressions-Models zur Vorhersage fehlender Werte auf Grundlage der Werte anderer Merkmale einer Observation
  - z.B. Lineare Regression, Logistische Regression, kNN

Vorteile

- höhere Genauigkeit als nicht-modell basierte Verfahren
- Abbildung nicht-linearer Zusammenhänge möglich
- Parallele Anwendung auf mehrere Merkmale

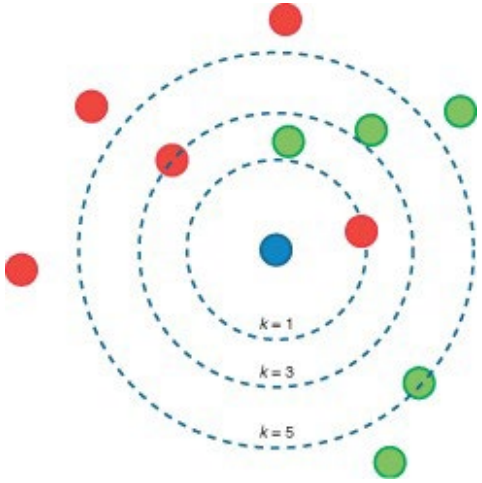
Nachteile

- Nur anwendbar, wenn ein tatsächlicher Zusammenhang zw. den Merkmalen besteht
- signifikant höhere Komplexität und Rechen-aufwand
- Einstellung von Hyperparametern erforderlich

Beispiel für modell-basierte Behandlung von Missing Values

k – Nearest Neighbours (kNN)

- **Idee:** „Missing Values“ werden aus den k Merkmalsausprägungen benachbarter Observationen geschätzt
- Anwendbar auf alle Merkmalsskalen
- Erfordert einen Zusammenhang zwischen dem Merkmal des „Missing Value“ und den anderen Merkmalen
- Das Finden benachbarter Observationen erfolgt über verschiedene Distanzmaße:



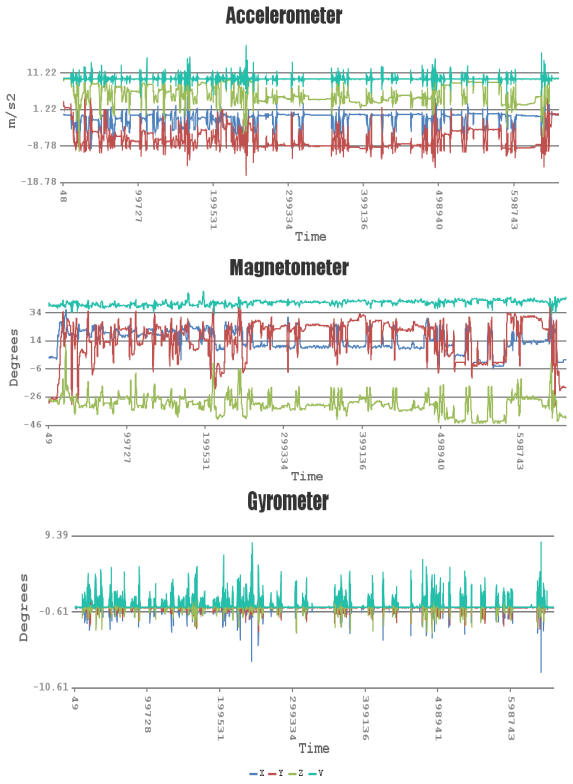
https://en.proft.me/media/science/r_knn_classify.png

Merkmalskalen	Abstandsmaß
Nominal / Ordinal	Hamming-Distanz, Gewichtete Hamming-Dist.
Metrisch	Euklidischer Abstand, Manhatten Distanz

Handling „Missing Values“

2. Behandlung von „Missing Values“

- Ersetzen in Zeitreihen:
  - **Last Observation Carried Forward (LOCF)**
    - Ersetzen aller fehlenden Merkmalsausprägungen einer Observation, durch die entsprechenden Ausprägungen der vorangehenden Observation
  - **Next Observation Carried Backward (NOCB)**
    - Ersetzen aller fehlenden Merkmalsausprägungen einer Observation, durch die entsprechenden Ausprägungen der nachfolgenden Observation
  - **Lineare Interpolation**
  - **Modus, Median, Mittelwert, Min oder Max der benachbarten n Observationen (je nach Skalierung der entsprechenden Merkmale)**
  - **Modellbasiert (ARIMA)**



https://fastercapital.com/de/thema/umgang-mit-ausrei%C3%9Fern-in-daten.html

3. Behandlung von Ausreißern

Behandlung von Ausreißern

Ausreißer

- Observationen, die von den Erwartungen (Verteilung) der zugrundeliegenden Merkmale signifikant abweichen

- **Ursachen:**
  - Menschlichte Fehler bei der Dateneingabe-/erfassung
  - Messfehler (bedingt durch fehlerhafte Messinstrumente)
  - Fehler im Versuchsaufbau
  - Bewusste Falschangaben durch Testpersonen
  - Fehler in der Vorverarbeitung
  - Natürliche Ausreißer („Novelties“)
  - Bewusst eingefügt für Test der Ausreißererkennung

Behandlung von Ausreißern

Ausreißer

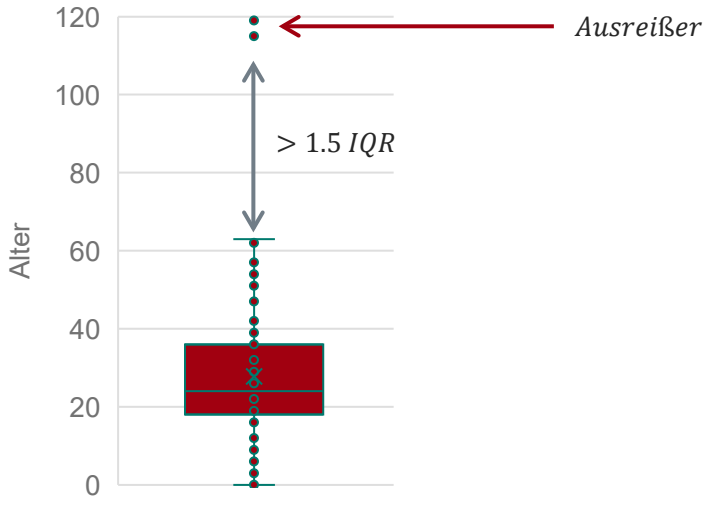
- Observationen, die von den Erwartungen (Verteilung) der zugrundeliegenden Merkmale signifikant abweichen

- **Auswirkungen:**
  - Einfluss auf statistische Kenngrößen (Mittelwert, Varianz, usw.)
  - Ggf. Einfluss auf die Genauigkeit der statistischen Modellierung der Daten
  - Ausreißer (Anomalien) sind nicht zwangsläufig als Fehler anzusehen, sollten jedoch im Rahmen der „Data Exploration“ untersucht werden

- **Spezielle Anwendungsgebiete der Ausreißererkennung**
  - Fraud detection (z.B. Erkennen einer Veränderung des Kaufverhalten nach dem Diebstahl einer Kreditkarte)
  - Predictive Maintainance (z.B. frühzeitige Erkennung eines Maschinendefekts auf Grundlage der Ausreißerraten in den Sensordaten)

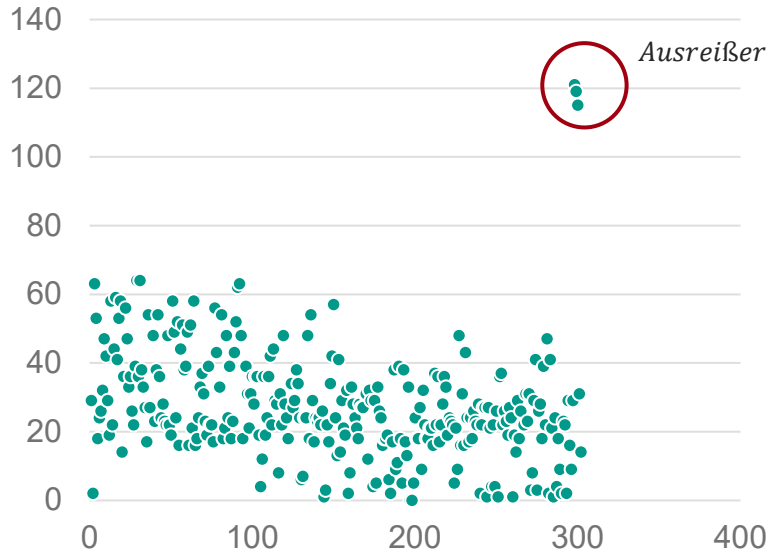
Identifikation von Ausreißern

Univariate Ausreißer



- Verwendung einfacher univariater statistischer Methoden / Visualisierung zur Erkennung von extremen Merkmalsausprägungen:
  - Boxplot, Histogramm
  - IQR-Distanz, z-Score

Multivariate Ausreißer



- Verwendung multivariater statistischer Methoden / Visualisierung zur Erkennung von extremen Merkmalsausprägungen (z.B: Scatterplot)
- Clusterbasierte Ansätze
- Dichtebasierte Ansätze
- Distanzbasierte Ansätze

Identifikation von Ausreißern (Multivariat)

Cluster-basierte Methoden

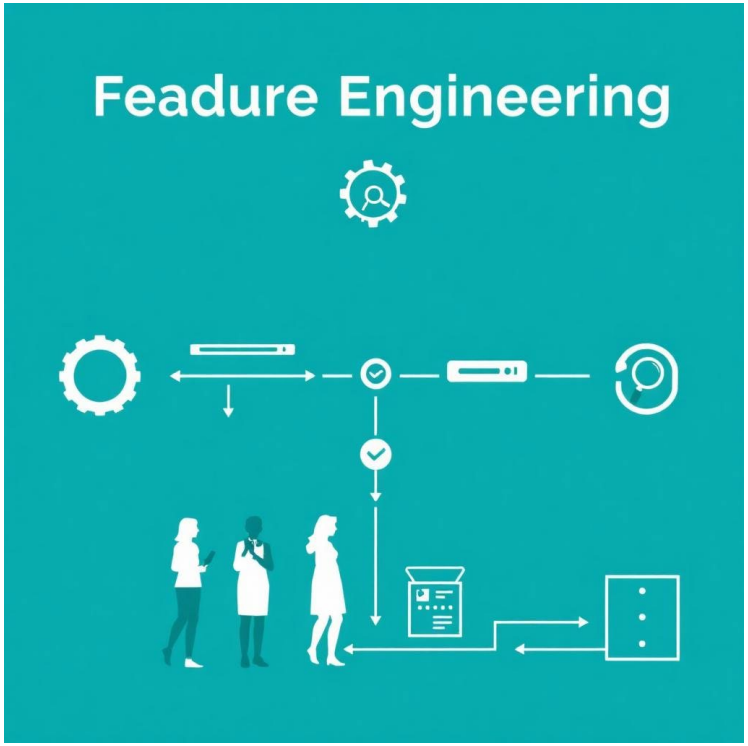
- Clusteranalyse: Datenpunkte werden zu n Clustern zusammengefasst
- Datenpunkte die keinem der n Cluster angehören sind Ausreißer
- z.B. k-means

Distanz-basierte Methoden

- Bewertet jeden Datenpunkt hinsichtlich seiner aggregierten Distanz zu seinen nächsten Nachbarn(kNN)
- Datenpunkte mit einem Distanzwert signifikant über dem Durchschnitt sind Ausreißer

Dichte-basierte Methoden

- Vergleich der Dichtedifferenzen zwischen Datenpunkten
- Annahme: Dichte ist nicht in jedem Bereich des Merkmalsraums gleich
- Ausreißer sind Datenpunkte, deren Dichte sich signifikant von der Dichte der benachbarten Datenpunkte unterscheidet
- z.B. Local Outlier Factor

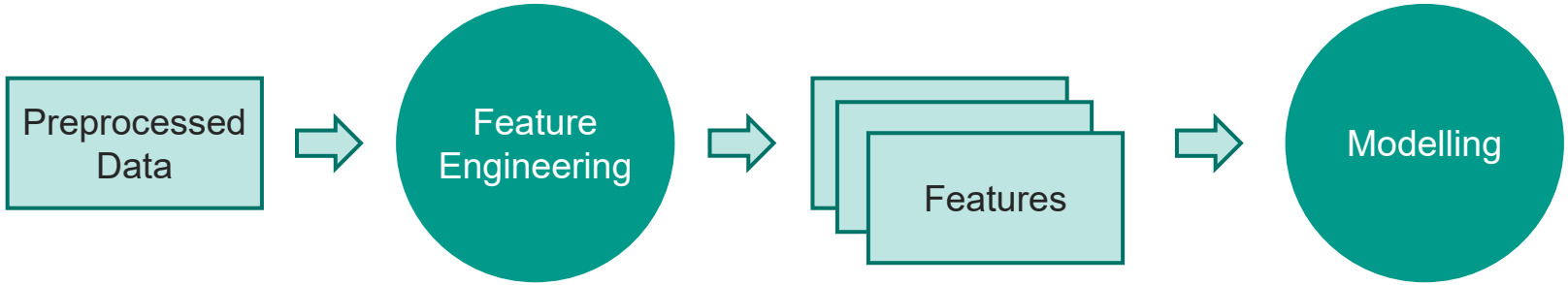


4. Feature Engineering

Feature Engineering

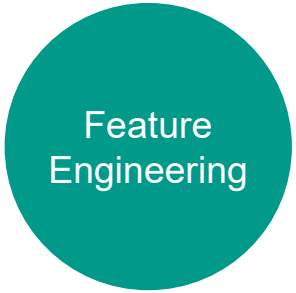
> "Coming up with features is difficult, time-consuming, requires expert knowledge. 'Applied machine learning' is basically feature engineering."
>
> Andrew Ng

> "Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data."
>
> Jason Brownlee

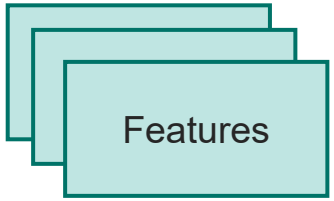


Feature Engineering

- Transformation der Ausgangsdaten in ein neues Format, das von Machine Learning Modellen besser interpretiert werden kann
- Reduziert die Komplexität und beschleunigt die Konvergenz der verwendeten Modelle
- Erfordert Expertenwissen über die zugrundeliegende Problemstellung
- Wird zunehmend automatisiert (z.B: durch Deep Learning)



- Repräsentieren spezielle problembezogene Charakteristiken der Ausgangsdaten (z.B. Extraktion lokaler Merkmale in Bilddaten)
- Schnittstelle zwischen den Daten und dem Model
- Beeinflussen maßgeblich die Genauigkeit des später trainierten Modells
- Sollten informativer für das Modell sein als die zugrundeliegenden Rohdaten



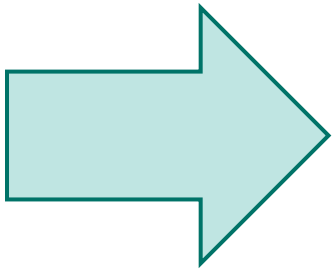
Feature Engineering

- Eine Vielzahl von Machine Learning Verfahren unterstützen ausschließlich metrisch skalierte Merkmale (bzw. Modelle sind performanter mit metrisch-skalierten Daten)
- Es gibt Ausnahmen: z.B. Entscheidungsbäume, Random Forest
- In der Regel ist ein Encoding nicht-metrischer Variablen sinnvoll
  - **Label Encoding**
  - **One Hot Encoding**
- Extraktion von (Teil-) Informationen aus Merkmalen oder Zusammenführen mehrerer Merkmale
  - **Extraction**
  - **Interactions**
- Bei Daten mit schiefen Verteilungen bietet sich ggf. eine Transformation an
  - **Log-Transformation**
- Metrische Merkmale unterschiedlicher Wertebereiche sollten auf einen Wertebereich skaliert werden
  - **Min-Max-Scaling**
  - **Mean-Normalisation**
  - **Standardisierung**

Label Encoding

- Jeder Ausprägung wird durch einen fixen metrischen Wert ersetzt
- Nur anwendbar auf Ordinale Merkmale
- Die Anwendung auf Nominale Merkmale würde eine nicht existente Ordnung der Werte erzeugen, die vom Model fälschlicherweise berücksichtigt werden könnte.

Class
First
Second
First
Third
Third

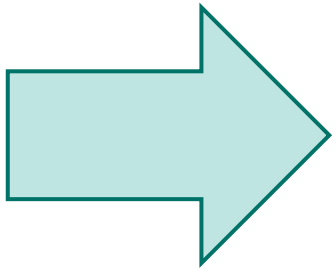


Class
1
2
1
3
3

One-Hot Encoding

- Jeder Ausprägung wird durch ein separates Merkmal abgebildet
- Anwendbar auf ordinale und nominale Merkmale
- Bei ordinalen Merkmalen geht die Ordnung durch das Encoding verloren

Embark
Southampton
Cherbourg
Southampton
Queenstown
Queenstown

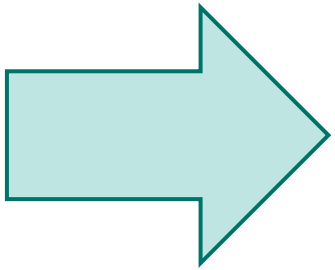


Sout-hampton	Cher-bourg	Queens-town
1	0	0
0	1	0
1	0	0
0	0	1
0	0	1

Extraction

- Ausprägungen nominaler/ordinal skalierter Merkmale enthalten unter Umständen relevante Informationen, die in ein separates Merkmal ausgelagert werden sollten.
- Bringt einen Verlust an Informationen mit sich (Einzelfallentscheidung, ob akzeptabel oder nicht)
- Beispiel: Extraktion des Stockwerks aus der Kabinennummer

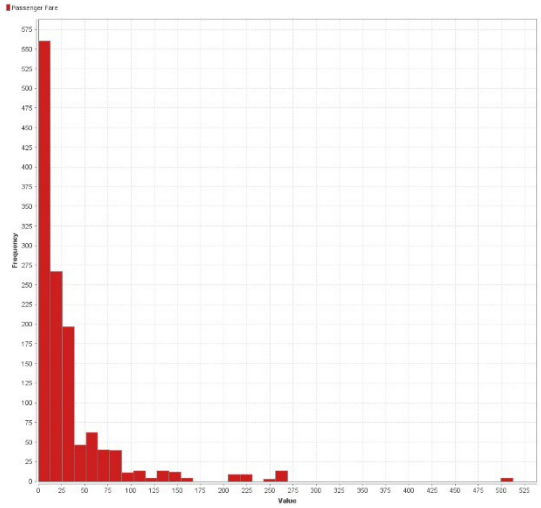
Cabin
A37
D7
T38
E2
A5



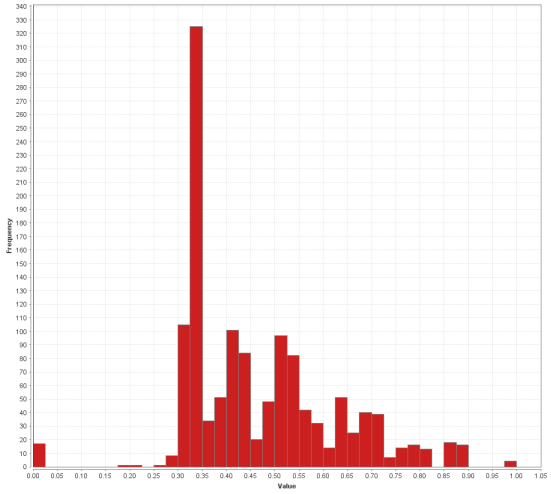
Floor
A
D
T
E
A

Log-Transformation

- Bei Merkmalen mit stark schiefen Verteilungen kann eine Transformation sinnvoll sein, um
  - eine Verteilung zu erzeugen, die sich der Normalverteilung annähert
  - die Größe des Merkmalsraums zu reduzieren
- Sollte mit Bedacht eingesetzt und validiert werden, da nicht immer zuverlässig (siehe Artikel)



x

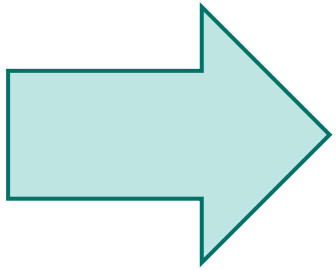


$\log(x)$

Binarization

- Ausprägungen eines metrischen Merkmals werden nach einer geeigneten Methode (z.B. Schwellwert) in zwei Klassen eingeteilt (Diskretisieren)
- Anwendbar auf ordinale und metrische Merkmale
- Bringt einen Verlust an Informationen mit sich (Einzelfallentscheidung, ob akzeptabel oder nicht)

Price
500
100
0
200
0

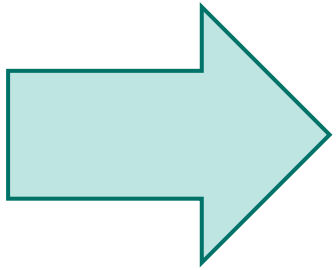


Free Ticket
0
0
1
0
1

Binning

- Ausprägungen eines metrischen Merkmals werden nach einer geeigneten Methode (Feste Schwellwerte, Quantile) in mehrere Klassen eingeteilt (Diskretisieren)
- Anwendbar auf ordinale und metrische Merkmale
- Bringt einen Verlust an Informationen mit sich (Einzelfallentscheidung, ob akzeptabel oder nicht)

Price
500
100
0
200
0



Price Cat.
2
1
0
1
0

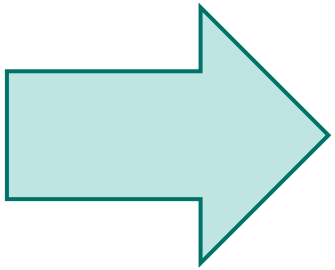
Scaling

- Viele Machine Learning Modelle gewichten Merkmale mit großen Ausprägungen stärker als Merkmale mit kleinen Ausprägungen
- Metrische Merkmale unterschiedlicher Wertebereiche sollten deshalb auf einen gemeinsamen Wertebereich skaliert werden (z.B: 0-1)
- **Min-Max-Scaling:**

Price	Age
5000	2
1000	68
0	12
2000	5
0	31

x

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$



$$0 < x_{scaled} < 1$$

Price	Age
1	0
0.2	1
0	0.152
0.4	0.045
0	0.439

x_{scaled}

Scaling

- Viele Machine Learning Modelle gewichten Merkmale mit großen Ausprägungen stärker als Merkmale mit kleinen Ausprägungen
- Metrische Merkmale unterschiedlicher Wertebereiche sollten deshalb auf einen gemeinsamen Wertebereich skaliert werden
- **Standardisation:**

Price	Age
5000	2
1000	68
0	12
2000	5
0	31

x

$$x_{scaled} = \frac{x - \bar{x}}{\sigma}$$

$\bar{x}_{scaled} = 0$ $\quad \bar{\sigma}_{scaled} = 1$

Price	Age
1,640	-0,792
-0,289	1,629
-0,772	-0,425
0,193	-0,682
-0,772	0,271

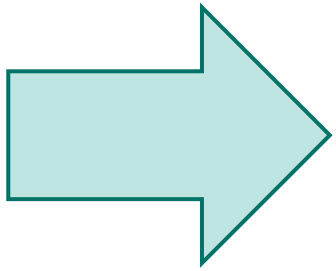
x_{scaled}

Feature Engineering (nominal/ordinal)

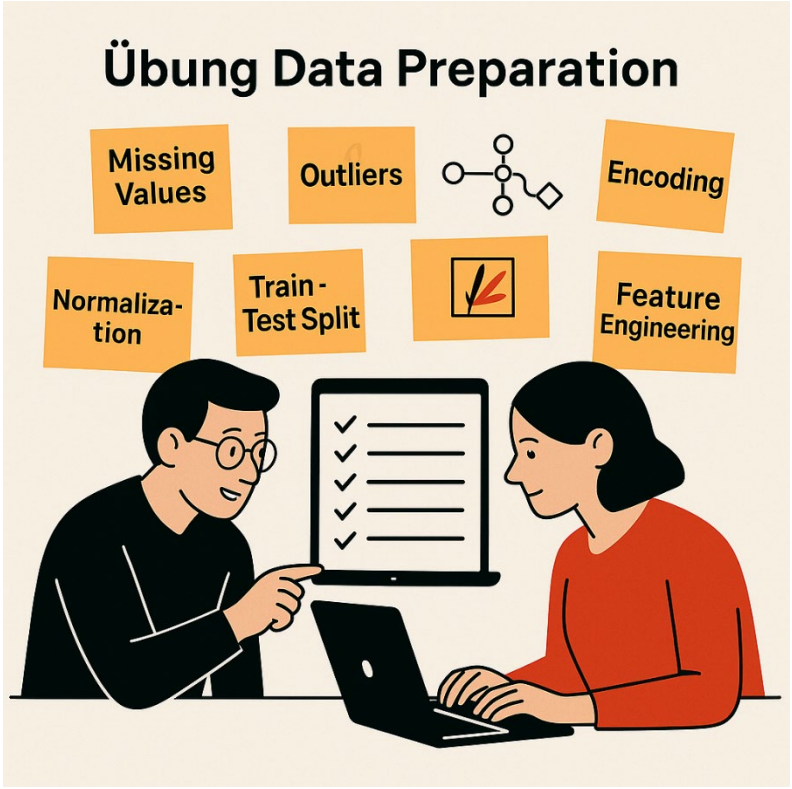
Interactions

- Ausprägungen verschiedener Merkmale lassen sich ggf. zu Ausprägungen eines neuen Merkmals zusammenfassen (z.B. durch Addition, Subtraktion bestehender Merkmale)
- Bringt einen Verlust an Informationen mit sich (Einzelfallentscheidung, ob akzeptabel oder nicht)
- Beispiel: Zusammenfassen der Merkmale „Anzahl Eltern/Kinder“ und „Anzahl Ehepartner/Geschwister“ zu einem neuen Merkmal „Anzahl Verwandte“

Child./Par.	Spou./Sib
1	0
0	0
0	3
4	2
2	1



Relatives
1
0
3
6
3


Übung Data Preparation
Missing Values
Outliers
Encoding
Normaliza-tion
Train - Test Split
Feature Engineering

5. Übung

Titanic Datensatz – Gruppenarbeit – Iris Challenge

Aufgabe

- Bilden Sie 2-er Gruppen.
- Untersuchen Sie die ausgeteilten Iris-Daten auf Änderungen (fehlende Werte, unplausible Werte, Duplikate etc.).

Vorgehen

- Vergleich mit Originaldatensatz nicht erlaubt!
- Überprüfen und korrigieren Sie den Datensatz mit den genannten Verfahren!
- Erstellen Sie dazu eine entsprechende Dokumentation
- Erläutern Sie Ihre Herangehensweise und die durchgeführten Änderungen
- Erkennt ChatGPT o.ä. die Änderungen?

Die Gruppe, die die meisten Veränderungen entdeckt und korrigiert, gewinnt die Challenge.