Datenvorbereitung

Teil A. Begriffe & Mathe Grundlagen

A1. Datensätze, Variablen, Typen

- Datensatz: Tabelle mit „n" Zeilen (Beobachtungen) und „d" Spalten (Merkmale / Features).

	0	1	0	...	0
	KPI	E_1	E_2	...	E_n
t_0	=	=	=	=	=
t_1		N.A.			
...					

- Numerische Merkmale: reelle Werte (zB. Preis, Größe)
- kategorische Merkmale: endliche Ausprägungen
  zB: Schuhe A, B, C, D.
- Ordinale Merkmale: (z.B. Schulnoten)
- Zeitstempel: Zeitpunkte/Zeiträume

A2. Fehlende Daten

- Darstellung: Ein Eintrag kann fehlen.
  Wir schreiben dazu „NA"
- Indikatorfunktion: Für Spalte „x" ist $M_i = 1$, wenn „x" fehlt; sonst $M_i = 0$.

· Arten von N.A. (Intuitive Beschreibung)

☐ MCAR. Missing Completely at Random
Das fehlen der Daten hat keinen Zusammenhang mit den Daten.

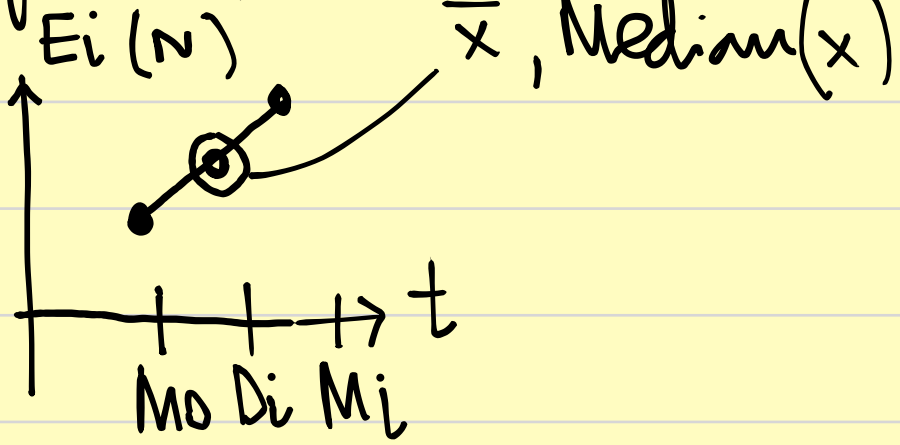
☐ MAR. Missing at Random
Fehlen hängt von beobachteten Variablen ab (zB. Information über Patienten wird ausgeblendet).

☐ MNAR. Missing not at Random.
Fehlen hängt mit dem fehlenden Wert selbst Zusammen

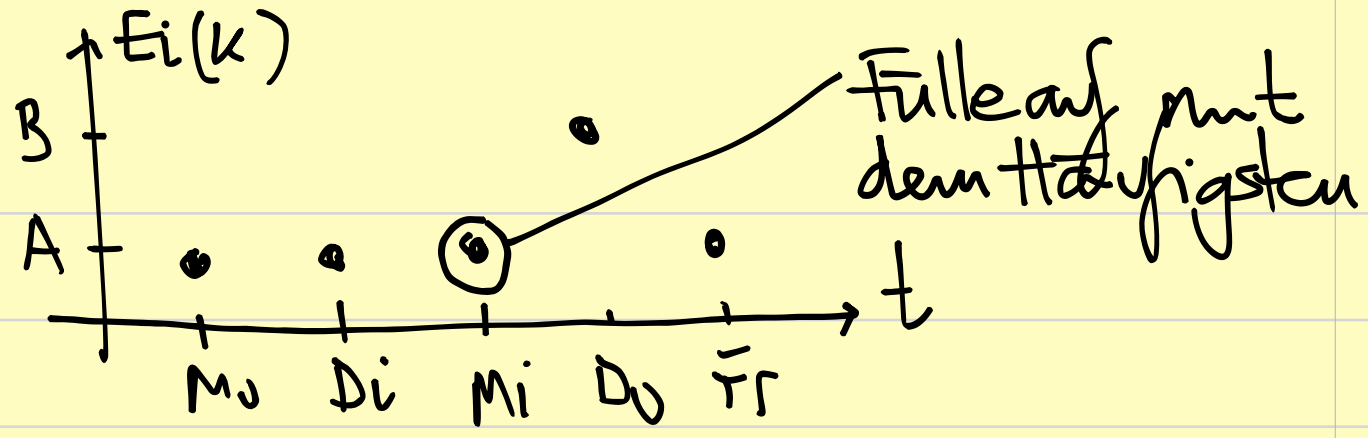
GOODHART's LAW.

· Imputation. (Auffüllen)

☐ Numerisch



☐ Kategorisch



. Fehlindikator als neue Feature, damit das Model „weiß", dass etwas gefehlt hat.

A3. Skalierung/Normierung numerische Merkmale

· Standardisierung (z-Score) $z_i = \frac{x_i - \bar{x}}{Std(x)}$

· Min. Max Normierung [0,1] (Normalisierung) $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$

· Robuste Skalierung $x' = \frac{x - med(x)}{IQR(x)}$

$IQR(x) = Q_{0'75} - Q_{0'25}$

unempfindlicher gegen Ausreißer.

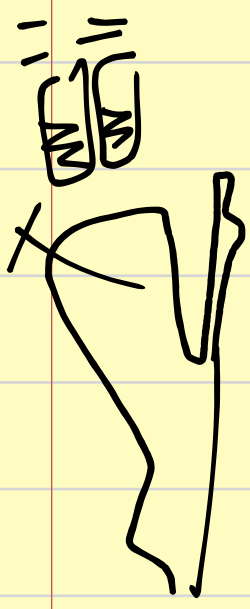
A4. Ausreißer (Outliers)

· IQR(Regel): Werte außerhalb dem Interval

$$\left[Q_{0'25} - 1'5\,IQR \;,\; Q_{0'75} + 1'5\,IQR \right]$$

gelten als Ausreißer.

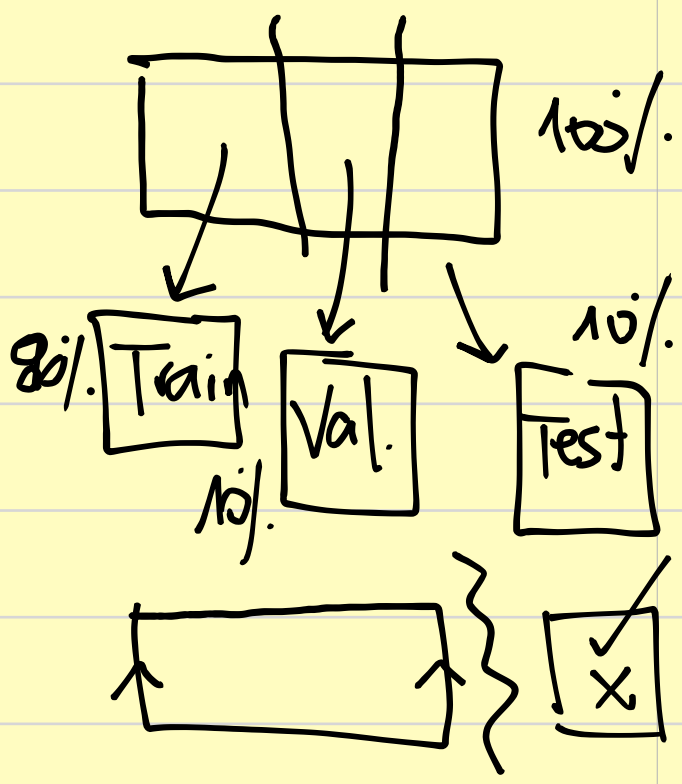
. Quantil-Kappen (Winsorizing): schneide extreme Werte auf z.B. 1% - 99% Quantil zurück.



WICHTIG: Kappen/Skalieren NUR bei Training. Bei Test/Validation → keine Datenleckage!

A5. Train/Validation/Test

80% 10% 10%



A6. Duplikate/Konsistenz/Einheiten

- Duplikate: identische Zeilen entfernen.
- Konsistente Kategorien: Groß/Kleinschreibung (zB Berlin/berlin/bIrlen)...
- Einheiten. (zB cm → m) Alle einheitlich darstellen.