Wörterliste ML Datenaufbereitung

Instanzen	- Einzelne Datenpunkte oder «Fälle», die das System lernen soll – z.B. ein Kunde bei TeleFlow - Herausforderungen: Daten über Instanzen kommen oft aus mehreren Quellen und müssen mittels Joins zusammengeführt werden.

Klassenattribut	- Das Ziel, das vorhergesagt werden soll (z.B. «Kunde kündigt Vertrag: Ja oder Nein»)
	- Herausforderungen: Muss teilweise manuell erstellt oder berechnet werden
	  o Manuell: z.B. bei der Frage, ob eine E-Mail «interessant» ist.
	  o Komplex: z.B. ob ein Kunde kündigt → aus seinem Verhalten abgeleitet

Attribute	- Merkmale/Beschreibungen einer Instanz (z.B. Wohnort, Anzahl, Tickets, letzte Reperatur)
	- Typische Aufgaben:
	  o Transformation: z.B. Datum in «Saison» umwandeln
	  o Feature Creation: neue Attribute aus bestehenden erstellen.
	  o Aggregation: historische Daten zusammenfassen.
	  o N:m-Problem: Einer Instanz können mehrere Werte eines Attributs zugeordnet sein
	  o Feature Selection: irrelevante Attribute entfernen (z.B: IDs, Datumsangaben)

Fehlende Werte	Umgangsmöglichkeiten: - Ignorieren: Sonderwert einfügen - Löschen 1: Instanzen mit fehlenden Attributwerten löschen(Zeile) - Löschen 2: Attribut mit fehlenden Werten aus allen Datenobjekte entfernen (Spalte) - Berechnen(«impute») 1: Mittelwert/Median/Modus einfügen - Berechnen(«impute») 2: Wert aus ähnlichster Instanz des Datensatzes kopieren

Cleaning (Datenbereinigung)	- Beseitigung von:
	  o Duplikaten
	  o Ausreissern
	  o Inkonsistenzen

Typkonversion	- Einfache Konversion: z.B. Monat «3» als Kategorie behandeln, nicht als Zahl.

<table>
<tr>
<td></td>
<td>
- Diskretisierung: numerische Werte in Kategorien einteilen (z.B. «jung», «alt»)
- One-Hot-Encoding: Kategorien in Zahlenvektoren umwandeln:

Vor One-Hot-Encoding:
<table>
<tr><th>Nahrungsmittel</th><th>Kalorien</th></tr>
<tr><td>Apple</td><td>95</td></tr>
<tr><td>Chicken</td><td>231</td></tr>
<tr><td>Broccoli</td><td>50</td></tr>
</table>

Nach One-Hot-Encoding:
<table>
<tr><th>Apple</th><th>Chicken</th><th>Broccoli</th><th>Calories</th></tr>
<tr><td>1</td><td>0</td><td>0</td><td>95</td></tr>
<tr><td>0</td><td>1</td><td>0</td><td>231</td></tr>
<tr><td>0</td><td>0</td><td>1</td><td>50</td></tr>
</table>
</td>
</tr>
<tr>
<td>Normalisierung</td>
<td>
- Alle Werte in einen Bereich zwischen 0 und 1 bringen
- Formel:
$$x_{neu} = \frac{x_{alt} - min}{max - min}$$
</td>
</tr>
<tr>
<td>Skalierung</td>
<td>
- Alle Werte so anpassen, dass sie:
  ○ einen Mittelwert von 0 haben.
  ○ und eine Standardabweichung von 1.
- Formel:
$$x_{neu} = \frac{x_{alt} - \bar{x}}{\sigma}$$
</td>
</tr>
</table>