

## Klausurvorbereitung

### Aufgabe 1

Gegeben sind folgende Daten:

Verkaufsdaten			Kunde-Kundenberater-Zuordnung			
Kunde	Periode	Umsatz	Kunden	Kundenberater	gültig von	gültig bis
K1	2017-04	250	K1	KB Y	01.01.1900	31.12.2015
K2	2017-04	50	K1	KB X	01.01.2016	31.12.2050
K3	2017-04	150	K2	KB X	01.01.1900	30.04.2017
K3	2017-04	100	K2	KB Y	01.05.2017	31.12.2050
K1	2017-05	200	K3	KB Y	01.01.1900	31.12.2050
K5	2017-05	150	K4	KB Y	01.01.1900	31.12.2050
K2	2017-05	100	K5	KB Y	01.05.2017	31.12.2017
K3	2017-05	50	K5	KB X	01.01.2018	31.12.2050
K4	2017-05	100				
K2	2017-05	100				
K5	2017-06	200				
K3	2017-06	50				
K2	2017-06	100				

Stellen Sie die monatlichen Umsätze der Kundenberater nach aktueller Struktur (Juni 2017), nach gültiger Struktur im April 2017, nach historischer Wahrheit und nach vergleichbaren Resultaten dar.

### Aufgabe 2

Gegeben ist folgende Tabelle:

Bestellinformationen	
Feld	Beschreibung
Bestell-ID	Eindeutige ID der Bestellung
Bestell-Datum	Datum der Bestellung
Lfd-Nr.	Laufende Nummer bezogen auf eine Bestell-ID
Kundenr.	Kundennummer (identifiziert einen Kunden eindeutig)
Kundenname	Name des Kunden
Kunde-Straße	Straße und Hausnummer, in der der Kunde wohnt
Kunde-PLZ	Postleitzahl des Wohnortes des Kunden
Kunde-Ort	Wohnort des Kunden
Produktnr.	Produktnummer (identifiziert ein Produkt eindeutig)
Produktname	Name des Produktes
Menge	Verkaufte Menge des zugeordneten Produktes bezogen auf die aktuelle Bestellung
Preis	Einzelpreis des zugeordneten Produktes

Der Primärschlüssel der Tabelle wird durch die Bestell-ID und die Lfd-Nr. gebildet.

Erstellen Sie mit Hilfe der zur Verfügung stehenden Informationen ein Stern-Schema nach den Regeln der dimensionalen Modellierung, mit dessen Hilfe u. a. folgende Fragestellungen beantwortet werden können:

- Wie hoch war der Umsatz von Produkt X im Mai 2017?
- In welchen PLZ-Gebieten wurden 2016 die meisten Produkte verkauft?
- In welchem Quartal werden in der Regel die höchsten Umsätze erzielt?
- Wer war im April 2017 der Top-Kunde bezogen auf die Anzahl gekaufter Produkte?

Stellen Sie Ihr Ergebnis grafisch im Sternformat dar, sodass Fakten- und Dimensionstabelle sowie Primär- und Fremdschlüssel der Tabellen erkennbar sind.

### Aufgabe 3

Gegeben sei folgende Datenbasis mit Warenkörben:

TID	Items
1	Äpfel, Clementinen, Erdbeeren
2	Äpfel, Erdbeeren, Grapefruits
3	Äpfel, Clementinen, Erdbeeren, Feigen, Grapefruits
4	Äpfel, Clementinen, Himbeeren
5	Birnen, Clementinen, Erdbeeren, Feigen, Grapefruits
6	Birnen, Clementinen, Erdbeeren, Grapefruits, Himbeeren

- Führen Sie auf Grundlage der Datenbasis und einem minSupport von 0,6 den ersten Schritt des Apriori-Algorithmus (*Finden häufiger Item-Mengen*) durch.
- Führen Sie auf Grundlage Ihres Ergebnisses, der Datenbasis und einer minKonfidenz von 0,8 den zweiten Schritt des Apriori-Algorithmus (*Generierung von Assoziationsregeln mit hoher Konfidenz*) durch.
- Berechnen Sie den Lift der gefundenen Assoziationsregeln.
- Welche Regeln sind – entsprechend ihren Lift-Werten – „interessant“?

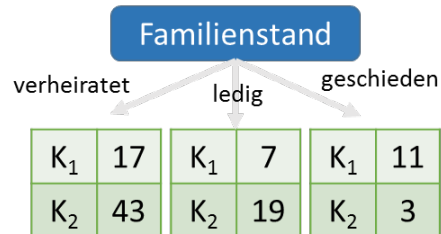
### Aufgabe 4

- Ein Klassifikationsmodell soll anhand des Kaufverhaltens prognostizieren, ob Kunden eines Online-Shops berufstätig, erwerbslos oder pensioniert sind. Das Modell wurde anhand historischer Daten getestet. Ergebnis ist untenstehende Konfusionsmatrix.

Konfusionsmatrix		Vorhergesagte Klasse		
		berufstätig	erwerbslos	pensioniert
Tatsächliche Klasse	berufstätig	143	8	6
	erwerbslos	3	12	1
	pensioniert	5	5	17

Berechnen Sie die Accuracy (Treffgenauigkeit) und die Error rate (Klassifikationsfehler) des Modells.

- b) Bei einem Entscheidungsbaum soll eine Menge von Kundendatenobjekten anhand des Attributs „Familienstand“ (mit den Ausprägungen „verheiratet“, „ledig“ und „geschieden“) aufgeteilt werden. Die Datenobjekte werden dabei den beiden Klassen  $K_1$  und  $K_2$  zugeordnet (siehe unten). Berechnen Sie für die drei entstehenden Knoten jeweils die Entropie und den Gini-Index.



### Aufgabe 5

Für einen aus drei Dokumenten bestehenden Korpus wurde – nach Normalisierung, Stemming und Entfernen der Stoppwörter – folgende Dokument-Term-Matrix (DTM) erstellt:

DTM	geig	lern	schach	spiel	tennis	üben
d <sub>1</sub>	0	0	1	2	1	0
d <sub>2</sub>	1	0	0	1	0	3
d <sub>3</sub>	2	1	0	1	0	1

- Wie könnte Dokument  $d_2$  konkret aussehen? Geben Sie einen korrekten deutschen Satz an, für den sich genau die Werte der DTM ergeben.
- Berechnen Sie für alle Terme  $t$  und für alle Dokumente  $d$  die Termfrequency  $TF(t, d)$ . Verwenden Sie bei Ihrer Berechnung den Logarithmus zur Basis 10.
- Berechnen Sie für alle Terme  $t$  die inverse Dokumentenhäufigkeit  $IDF(t)$ . Nutzen Sie wieder den Logarithmus zur Basis 10.
- Stellen Sie die TF-IDF-Matrix (mit den von Gerard Salton vorgeschlagenen Gewichten  $w(t, d)$ ) auf.
- Welche beiden Dokumente sind auf Basis der TF-IDF-Matrix vermutlich am ähnlichsten? (Eine genaue Berechnung der Ähnlichkeit ist nicht notwendig.)