Aufgabe 1

Gegeben sind folgende Daten:

Verkaufsdaten

Kunde	Periode	Umsatz
K1	2017-04	250
K2	2017-04	50
K3	2017-04	150
K3	2017-04	100
K1	2017-05	200
K5	2017-05	150
K2	2017-05	100
K3	2017-05	50
K4	2017-05	100
K2	2017-05	100
K5	2017-06	200
K3	2017-06	50
K2	2017-06	100

Kunde-Kundenberater-Zuordnung

Kunden	Kundenberater	gültig von	gültig bis
K1	KB Y	01.01.1900	31.12.2015
K1	KB X	01.01.2016	31.12.2050
K2	KB X	01.01.1900	30.04.2017
K2	KB Y	01.05.2017	31.12.2050
K3	KB Y	01.01.1900	31.12.2050
K4	KB Y	01.01.1900	31.12.2050
K5	KB Y	01.05.2017	31.12.2017
K5	KB X	01.01.2018	31.12.2050

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

aktuelle Struktur

	2017-04	2017-05	2017-06
KB X	250	200	-
KB Y	300	500	350

Struktur im April 2017

	2017-04	2017-05	2017-06
KB X	300	400	100
KB Y	250	150	50

historische Wahrheit

	2017-04	2017-05	2017-06
KB X	300	200	-
KB Y	250	500	350

vergleichbare Resultate

	2017-04	2017-05	2017-06
KB X	250	200	-
KB Y	250	150	50

Aufgabe 2

Gegeben ist folgende Tabelle:

Bestellinformationen

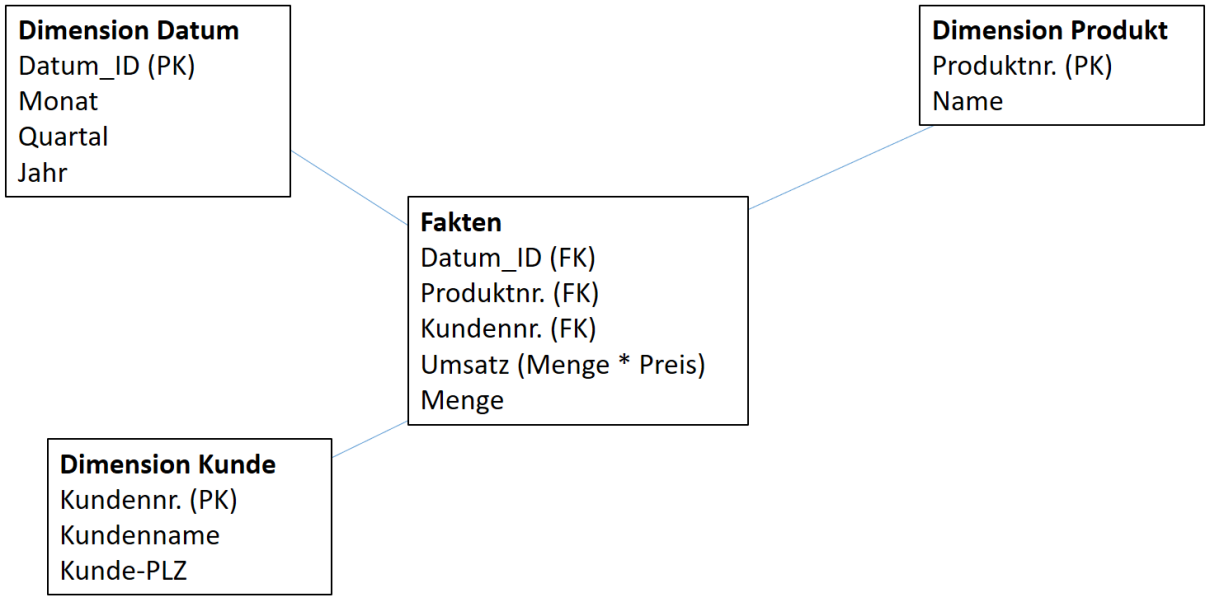
Feld	**Beschreibung**
Bestell-ID	Eindeutige ID der Bestellung
Bestell-Datum	Datum der Bestellung
Lfd-Nr.	Laufende Nummer bezogen auf eine Bestell-ID
Kundenr.	Kundennummer (identifiziert einen Kunden eindeutig)
Kundenname	Name des Kunden
Kunde-Straße	Straße und Hausnummer, in der der Kunde wohnt
Kunde-PLZ	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.


Dimension Datum
Datum_ID (PK)
Monat
Quartal
Jahr
Dimension Produkt
Produktnr. (PK)
Name
Fakten
Datum_ID (FK)
Produktnr. (FK)
Kundennr. (FK)
Umsatz (Menge * Preis)
Menge
Dimension Kunde
Kundennr. (PK)
Kundenname
Kunde-PLZ

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

a) Führen Sie auf Grundlage der Datenbasis und einem minSupport von 0,6 den ersten Schritt des Apriori-Algorithmus (*Finden häufiger Item-Mengen*) durch.

Items (k = 1)	**Support**
{Äpfel}	4/6 = 0,67
~~{Birnen}~~	~~2/6 = 0,33~~
{Clementinen}	5/6 = 0,83
{Erdbeeren}	5/6 = 0,83
~~{Feigen}~~	~~2/6 = 0,33~~
{Grapefruits}	4/6 = 0,67
~~{Himbeeren}~~	~~2/6 = 0,33~~

Items (k = 2)	**Support**
~~{Äpfel, Clementinen}~~	~~3/6 = 0,50~~
~~{Äpfel, Erdbeeren}~~	~~3/6 = 0,50~~
~~{Äpfel, Grapefruits}~~	~~2/6 = 0,33~~
{Clementinen, Erdbeeren}	4/6 = 0,67
~~{Clementinen, Grapefruits}~~	~~3/6 = 0,50~~
{Erdbeeren, Grapefruits}	4/6 = 0,67

Items (k = 3)	**Support**
~~{Clementinen, Erdbeeren, Grapefruits}~~	~~3/6 = 0,50~~

b) 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.

Regeln (k = 1)	**Konfidenz**
{Clementinen}→{Erdbeeren}	4/5 = 0,8
{Erdbeeren}→{Clementinen}	4/5 = 0,8
{Erdbeeren}→{Grapefruit}	4/5 = 0,8
{Grapefruit}→{Erdbeeren}	4/4 = 1,0

c) Berechnen Sie den Lift der gefundenen Assoziationsregeln.

Regeln	**Lift**
{Clementinen}→{Erdbeeren}	0,96
{Erdbeeren}→{Clementinen}	0,96
{Erdbeeren}→{Grapefruit}	1,2
{Grapefruit}→{Erdbeeren}	1,2

d) Welche Regeln sind – entsprechend ihren Lift-Werten – „interessant“?
Die Regeln {Erdbeeren}→{Grapefruit} und {Grapefruit}→{Erdbeeren} sind 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.

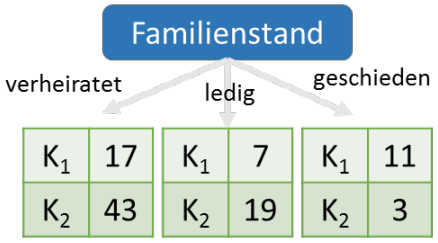
Konfusionsmatrix		**Vorhergesagte Klasse**		
		berufstätig	erwerbslos	pensioniert
Tatsächliche Klasse	berufstätig	143	8	6
	erwerbslos	3	12	1
	pensioniert	5	5	17

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

Accuracy = (143 + 12 + 17) / (143 + 8 + 6 + 3 + 12 + 1 + 5 + 5 + 17) = 0,86

Error rate = 1 – Accuracy = 0,14

a) 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.



Entropie(verheiratet) = - (17/60 * $\log_2$(17/60) + 43/60 * $\log_2$(43/60)) = 0,86
Entropie(ledig) = - (7/26 * $\log_2$(7/26) + 19/26 * $\log_2$(19/26)) = 0,8404
Entropie(geschieden) = - (11/14 * $\log_2$(11/14) + 3/14 * $\log_2$(3/14)) = 0,7496

Gini(verheiratet) = 1 - $(17/60)^2$ - $(43/6o)^2$ = 0,4061
Gini(ledig) = 1 - $(7/26)^2$ - $(19/26)^2$ = 0,3935
Gini(geschieden) = 1 - $(11/14)^2$ - $(3/14)^2$ = 0,3367

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_1	0	0	1	2	1	0
d_2	1	0	0	1	0	3
d_3	2	1	0	1	0	1

a) Wie könnte Dokument d_2 konkret aussehen? Geben Sie einen korrekten deutschen Satz an, für den sich genau die Werte der DTM ergeben.

Beispielsatz: „Will man Geige spielen, so muss man üben, üben und noch einmal üben!“

b) 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.

TF(t,d)	geig	lern	schach	spiel	tennis	Üben
d_1	0	0	1	1,30	1	0
d_2	1	0	0	1	0	1,48
d_3	1,30	1	0	1	0	1

c) Berechnen Sie für alle Terme *t* die inverse Dokumentenhäufigkeit *IDF(t)*. Nutzen Sie wieder den Logarithmus zur Basis 10.

	geig	lern	schach	spiel	tennis	Üben
IDF(t)	1,18	1,48	1,48	1	1,48	1,18

d) Stellen Sie die TF-IDF-Matrix (mit den von Gerard Salton vorgeschlagenen Gewichten *w(t, d)*) auf.

TF-IDF	geig	lern	schach	spiel	tennis	Üben
d_1	0	0	1,48	1,30	1,48	0
d_2	1,18	0	0	1	0	1,75
d_3	1,53	1,48	0	1	0	1,18

e) Welche beiden Dokumente sind auf Basis der TF-IDF-Matrix vermutlich am ähnlichsten? (Eine genaue Berechnung der Ähnlichkeit ist nicht notwendig.)

Die Dokumente d2 und d3 (Es gibt Teilübereinstimmungen in drei Dimensionen)