NAME DES DOZENTEN: NEUHAUS, SCHULZ, WELTER

KLAUSUR: I165 – ANALYTISCHE INFORMATIONSSYSTEME

QUARTAL: (2/2021)

Dauer: 90 Min

Datum: 2021-06-22

Seiten der Klausuraufgaben mit Deckblatt: 9

Hilfsmittel: NORDAKADEMIE-Taschenrechner

Bemerkungen:

- ***Bitte prüfen Sie zunächst die Klausur (alle Teile) auf Vollständigkeit***
- ***Bitte vermerken Sie auf Ihren Antwortbögen folgende Angaben:***
  - ***Name***
  - ***Matrikelnummer***
  - ***Zenturie***
  - ***Seitenzahl***

Es sind 100 Punkte erreichbar!
Zum Bestehen der Klausur sind 50 Punkte ausreichend!

Aufgabe	**Erreichbare Punkte**	**Erreichte Punkte**
Aufgabe 1: Historisierung	9	
Aufgabe 2: Stern-Schema	18	
Aufgabe 3: Assoziationsanalyse	19	
Aufgabe 4: Entscheidungsbaum	19	
Aufgabe 5: Text-Mining	6	
Aufgabe 6: Vorgehensmodell	15	
Aufgabe 7: Gemischtes	14	
Bonusaufgabe	4	
Summe	100+4	

Note:_____ Prozentsatz: _____ Ergänzungsprüfung:_____

Datum:_____ Unterschrift:_____

Datum:_____ Unterschrift:_____

Aufgabe 1: Historisierung (9 Punkte)

Gegeben sind folgende zwei Datenstände der Tabelle PRODUKT:

PRODUKT (Datenstand 1)

P_ID	P_BEZEICHNUNG	P_GEWICHT	P_GRUPPE
P001	Schokoriegel	125	Schokolade
P002	Pfefferminz-Kaugummi	50	Kaugummi
P003	Fruchtgummi	250	Gummibonbon
P004	Lakritz-Kaubonbon	200	Kaugummi

PRODUKT (Datenstand 2)

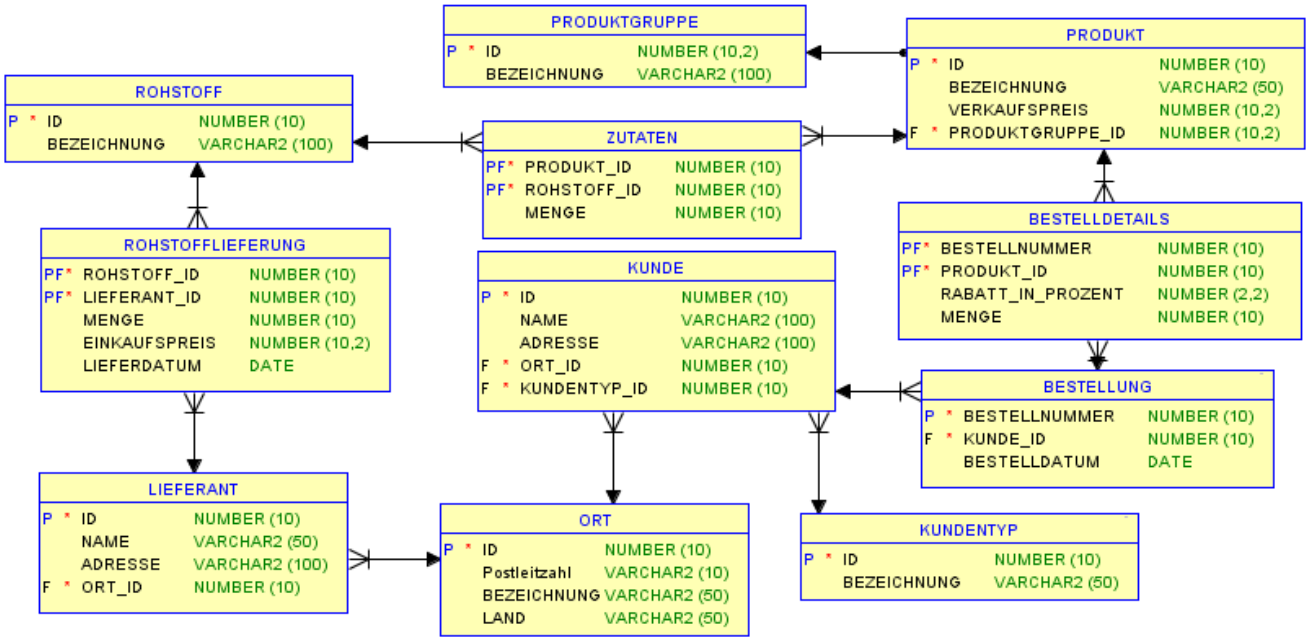
P_ID	P_BEZEICHNUNG	P_GEWICHT	P_GRUPPE
P002	Pfefferminz-Kaugummi	50	Kaugummi
P003	Fruchtgummi	250	Gummibonbon
P004	Lakritz-Kaubonbon	250	Gummibonbon
P005	Tafelschokolade	100	Schokolade

Das Attribut P_ID bildet den Primärschlüssel der Tabelle, P_GEWICHT und P_GRUPPE können sich ggf. ändern. Die Daten aus der Quelltabelle PRODUKT werden über einen ETL-Prozess in die Dimensionstabelle DIM_PRODUKT eingefügt, die nach den Prinzipien der Slowly Changing Dimension Typ 2 mit Tupelzeitstempelung historisiert wird. Am 01.01.2021 wurde der Datenstand 1 in die bis dahin leere Tabelle DIM_PRODUKT eingefügt. Am 01.06.2021 wurden die Daten in der Dimensionstabelle durch den Datenstand 2 aktualisiert. Stellen Sie alle Daten in der Tabelle DIM_PRODUKT (Spalten: P_ID, P_BEZEICHNUNG, P_GEWICHT, P_GRUPPE, GUELTIG_VON, GUELTIG_BIS) nach dem Laden beider Datenstände dar.

P_ID	P_Bezeichnung	P_Gewicht	P_Gruppe	Gültig_von	Gültig_bis
P001	Schokoriegel	125	Schoko	01.01.21	uc
P002	Pfefferminz-Kaug.	50	Kaug.	01.01.21	uc
P003	Fruchtg.	250	Gummibon.	01.01.21	uc
P004	Lakritz-Kaubon.	200	Kaug.	01.01.21	31.05.21
P004	Lakritz-Kaubon.	250	Gummibon.	01.06.21	uc
P005	Tafelschokolade	100	Schoko	01.06.21	uc

Aufgabe 2: Stern-Schema (18 Punkte)

Gegeben ist folgendes Datenmodell einer operativ verwendeten Datenbank eines Süßwarenherstellers.



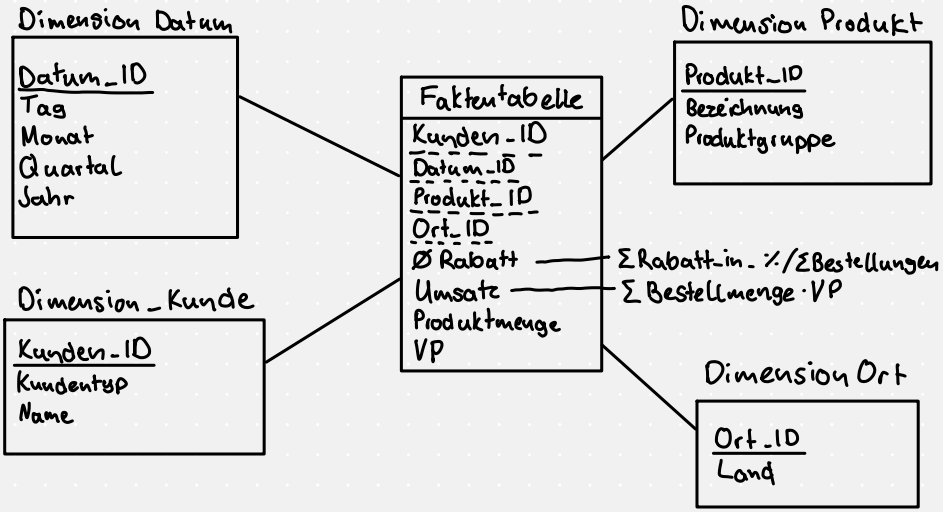
Lesehinweise: P – Primärschlüssel, F – Fremdschlüssel

a.) (15 Punkte) Erstellen Sie ein Stern-Schema nach den Regeln der dimensionalen Modellierung. Stellen Sie Ihr Ergebnis grafisch im Sternformat dar, sodass Fakten- und Dimensionstabelle sowie Primär- und Fremdschlüssel der Tabellen erkennbar sind. Sollten berechnete Kennzahlen nötig sein, geben Sie die Berechnungsregeln an. Sollten Sie Attribute umbenennen müssen, geben Sie die Namen der ursprünglichen Attribute und Tabellen an. Beschränken Sie das Stern-Schema auf die Inhalte, die benötigt werden, um folgende Fragen zu beantworten:

(1) Welchem Kundentyp wird der durchschnittlich höchste Rabatt gewährt?
(2) Welcher Umsatz wurde durch den Kunden „NORDAKADEMIE“ im Mai 2021 erzielt?
(3) Wie hoch war der Umsatz der Produktgruppe „Schokolade“ im vierten Quartal 2020?
(4) Wie viele Produkte wurden am 01.06.2021 insgesamt in Deutschland verkauft?
(5) Welches Produkt, das durch den Kunden „ACME“ bestellt wurde, hat den höchsten Verkaufspreis?

b.) (3 Punkte) Welches Problem würde entstehen, wenn zusätzlich zu den in a.) geforderten Fragen auch eine Analysemöglichkeit bezogen auf die Lieferanten in das Stern-Schema integriert werden sollte?

2. a)


Dimension Datum
Datum_ID
Tag
Monat
Quartal
Jahr
Dimension Produkt
Produkt_ID
Bezeichnung
Produktgruppe
Faktentabelle
Kunden_ID
Datum_ID
Produkt_ID
Ort_ID
Ø Rabatt
Umsatz
Produktmenge
VP
Σ Rabatt_in_% / Σ Bestellungen
Σ Bestellmenge · VP
Dimension_Kunde
Kunden_ID
Kundentyp
Name
Dimension Ort
Ort_ID
Land

Aufgabe 3: Assoziationsanalyse (19 Punkte)

Ein Online-Blumenversand erlaubt seit einiger Zeit die individuelle Zusammenstellung von Blumensträußen, also die persönliche Kombination von verschiedenen Blumenarten. Aufgrund der dabei gesammelten Daten soll nun bestimmt werden, welche Blumensorten häufig miteinander kombiniert werden. Ziel ist es, der Kundschaft beim Bestellprozess zu den bereits gewählten Blumenarten geeignete Ergänzungsvorschläge zu machen.

Folgende zehn individuelle Blumensträuße wurden im Online-Blumenversand bereits bestellt:

Blumenstrauß_ID	Chrysanthemen	Gerbera	Lilien	Nelken	Orchideen	Rosen	Tulpen
1	X	X			X		X
2		X		X	X		X
3	X	X			X		X
4		X	X		X	X	X
5		X				X	X
6			X	X			X
7	X	X			X	X	
8		X	X		X		X
9		X	X			X	X
10		X		X	X		X

a) Führen Sie auf Grundlage der Datenbasis und einem **minSupport** von **0,5** den ersten Schritt des Apriori-Algorithmus (*Finden häufiger Item-Mengen*) durch. *(5 Punkte)*
Hinweis: Sie können die Blumenarten durch ihren Anfangsbuchstaben abkürzen.

b) Führen Sie nun den zweiten Schritt des Apriori-Algorithmus (*Generierung von Assoziationsregeln mit hoher Konfidenz*) durch. Arbeiten Sie dabei mit einer **minKonfidenz** von **0,87**. *(6 Punkte)*

c) Berechnen Sie den Lift der hoch-konfidenten Assoziationsregeln. *(4 Punkte)*

d) Welche Regeln sind – entsprechend ihren Lift-Werten – „interessant"? Welche der Regeln halten Sie für am nützlichsten? Begründen Sie Ihre Antwort kurz. *(4 Punkte)*

Formeln zur Assoziationsanalyse

Items G: Grundgesamtheit G von Bezeichnern

Item-Menge X: Nicht-leere Teilmenge $X \subseteq G$

k-Item-Menge: Item-Menge mit k Elementen ($k \geq 1$)

Datenbank/-basis D: Menge D von Item-Mengen (Transaktionen); die Anzahl der Transaktionen ist $|D|$

Absolute Häufigkeit $n(X)$: Anzahl der Item-Menge X in der Datenbank D

Assoziationsregel: Implikation der Form $X \rightarrow Y$ mit $X \cap Y = \emptyset$

Support einer Item-Menge X: $Support(X) = \frac{n(X)}{|D|}$

Support einer Assoziationsregel $X \rightarrow Y$: $Support(X \rightarrow Y) = Support(X \cup Y)$

Konfidenz einer Assoziationsregel $X \rightarrow Y$: $Konfidenz(X \rightarrow Y) = \frac{n(X \cup Y)}{n(X)}$

Lift einer Assoziationsregel $X \rightarrow Y$: $Lift(X \rightarrow Y) = \frac{Konfidenz(X \rightarrow Y)}{Support(Y)} = \frac{\hat{P}(Y|X)}{\hat{P}(Y)}$

a)

Item k=1	Support
G	0,9
T	0,9
O	0,7
~~R~~	~~0,4~~
~~L~~	~~0,4~~
~~C~~	~~0,3~~
~~N~~	~~0,3~~

Item k=2	Support
G,T	0,8
G,O	0,7
T,O	0,6

Item k=3	Support
G,T,O	0,6

b)

Regel	Konfidenz
G→T	0,89
T→G	0,89
~~G→O~~	~~0,78~~
O→G	1
~~T→O~~	~~0,67~~
~~O→T~~	~~0,86~~
~~G,T→O~~	~~0,75~~
~~G,O→T~~	~~0,86~~
O,T→G	1
~~G→T,O~~	~~0,67~~
~~O→G,T~~	~~0,86~~
~~T→G,O~~	~~0,67~~

c)

Regel	Lift
G→T	0,99
T→G	0,99
O→G	1,11
O,T→G	1,11

d)

O→G

O,T→G

Aufgabe 4: Entscheidungsbaum (19 Punkte)

Die Dozenten des Moduls „Analytische Informationssysteme“ wollen vorhersagen, ob ihre Studierenden im ersten Versuch die Klausur bestehen. Grundlage dafür sind folgende Daten von Studierenden aus dem letzten Jahr:

Anwesenheit bei der Vorlesung	Beteiligung am Moodle-Diskussionsforum	Note im Modul „Algorithmen“	Klausur im ersten Versuch bestanden
meistens	Häufig	Zwei oder besser	ja
immer	Selten	Zwei oder besser	nein
gelegentlich	Selten	Drei oder schlechter	ja
meistens	Nie	Zwei oder besser	nein
gelegentlich	Häufig	Drei oder schlechter	ja
meistens	Selten	Zwei oder besser	ja
immer	Nie	Drei oder schlechter	nein
gelegentlich	Nie	Zwei oder besser	ja
meistens	Häufig	Drei oder schlechter	nein
immer	Häufig	Drei oder schlechter	ja

a.) Erstellen Sie dazu unter Verwendung des Hunt-Algorithmus und der Entropie einen Entscheidungsbaum. Splitten Sie dabei alle Attribute jeweils maximal auf. Geben Sie alle benötigten Rechenschritte und Entscheidungen an. Stoppen Sie den Algorithmus bereits nach der ersten Aufsplittung. *(10 Punkte)*

b.) Zeichnen Sie den resultierenden Entscheidungsbaum. *(4 Punkte)*

c.) Berechnen Sie für Ihren Entscheidungsbaum die Accuracy/Treffgenauigkeit. Skizzieren Sie dabei Ihren Rechenweg. (2 Punkte)

d.) Beschreiben Sie, wie der Algorithmus ohne Stopp-Kriterium fortfahren würde. (3 Punkte)

Entscheidungsbäume

Accuracy (Treffgenauigkeit): $\text{Accuracy} = \frac{korrekte\ Klassifikationen}{alle\ Klassifikationen} = \frac{r_p+r_n}{r_p+f_p+f_n+r_n}$

Error rate (Klassifikationsfehler): $\text{Error rate} = \frac{falsche\ Klassifikationen}{alle\ Klassifikationen} = \frac{f_p+f_n}{r_p+f_p+f_n+r_n}$

Homogenitätsmaße:

Gegeben: Knoten T mit $|T|$ Datensätzen in k Klassen (Partitionen von T)
p_i = relative Anzahl der Datensätze der Klasse i (mit $i = 1, \ldots, k$)

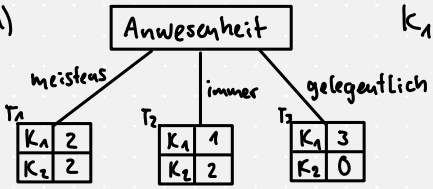
Entropy des Knotens T: $\text{Entropy}(T) = -\sum_{i=1}^{k} p_i \log_2 p_i$

Gini-Index des Knotens T: $\text{Gini}(T) = 1 - \sum_{i=1}^{k} p_i^2$

Klassifikationsfehler des Knotens T: $\text{Klassifizierungsfehler}(T) = 1 - \max\{p_i\}$

Information Gain bei Aufspaltung des Knotens T: $\text{Information Gain} = H(T) - \sum_{i=1}^{m} \frac{|T_i|}{|T|} H(T_i)$

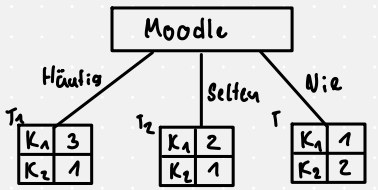
a)



K_1 = ja ; K_2 = nein $\quad G(T_2) = 1 - \left(\frac{1}{3}^2 + \frac{2}{3}^2\right) = \frac{4}{9}$

$Gini(T_1) = 1 - \left(\frac{2}{4}^2 + \frac{2}{4}^2\right) = \frac{1}{2} \quad G(T_3) = 1 - \frac{3}{3}^2 = 0$

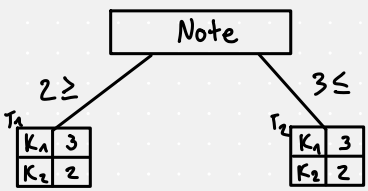
$Gini_{komplett} = \frac{4}{10} \cdot \frac{1}{2} + \frac{3}{10} \cdot \frac{4}{9} = \frac{1}{3}$



$G(T_1) = 1 - \left(\frac{3}{4}^2 + \frac{1}{4}^2\right) = \frac{3}{8}$

$G(T_2) = 1 - \left(\frac{2}{3}^2 + \frac{1}{3}^2\right) = \frac{4}{9} \quad Gini_{komp} = \frac{4}{10} \cdot \frac{3}{8} + \frac{3}{10} \cdot \frac{4}{9} + \frac{3}{10} \cdot \frac{4}{9} = \frac{5}{12}$

$G(T_3) = \frac{4}{9}$

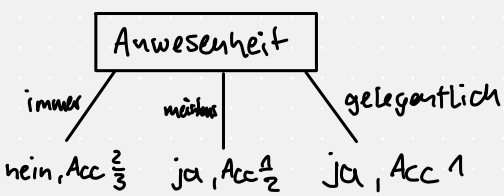


$G(T_1) = 1 - \left(\frac{3}{5}^2 + \frac{2}{5}^2\right) \approx 0{,}48$

$G(T_2) = 1 - \left(\frac{3}{5}^2 + \frac{2}{5}^2\right) \approx 0{,}48$

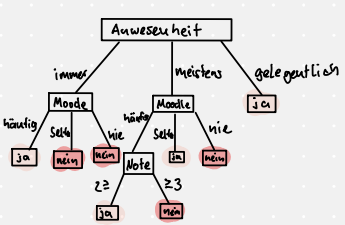
$Gini_{komp.} = \frac{5}{10} \cdot 0{,}48 \cdot 2 = 0{,}48$

b)



c) $\frac{2+2+3}{10} = \frac{7}{10} = 70\,\%$

d) weiter aufsplitten mit jeweils höchsten IG Knoten.
Acc wäre dann 100%.



Aufgabe 5: Text-Mining (6 Punkte)

Gegeben sei die folgende Term-Dokument-Matrix:

	essen	**apfel**	**hunger**
d_1	2	0	1
d_2	0	1	0
d_3	0	1	2

a) Berechnen Sie die Inverse Document Frequency (IDF) für *essen* und *apfel* mit zwei Nachkommastellen. Ihr Rechenweg muss erkennbar sein. Nutzen Sie den Zehnerlogarithmus. (2 Punkte)

$IDF(essen) = 1 + \log_{10}\left(\frac{3}{1}\right) \approx 1{,}48 \qquad IDF(apfel) = 1 + \log_{10}\left(\frac{3}{2}\right) \approx 1{,}18$

Gesamtzahl der Dokumente: N
Anzahl der Dokumente mit Term t: $df(t)$
Inverse Document Frequency: $IDF(t) = 1 + \log(\frac{N}{df(t)})$

b) Mit welchem Distanzmaß wird die Ähnlichkeit zwischen Dokumenten gemessen? Warum wird dafür nicht der euklidische Abstand verwendet? (2 Punkte)

c) Erläutern Sie anhand von zwei Beispielen des folgenden Satzes, wodurch ein Tokenizer Fehler produzieren kann. (2 Punkte)
„New York wurde im 20. Jh. zum Zentrum für Industrie und Handel. Am 24. Oktober 1929 jedoch begann eine schlimme Wirtschaftskrise."

b) Cosinus, weil man

c) Punkt hinter Zahl als Satzende

Aufgabe 6: Vorgehensmodell (15 Punkte)

Gegeben sei folgendes, kurzes Fallbeispiel. Die Nummern dienen nur zum Referenzieren der Sätze und sind nicht inhaltlich relevant.
Peter arbeitet als Data-Science-Consultant.
(1) Bei seinem aktuellen Auftrag hat er sich zu Beginn für zwei Tage mit den Auftraggebern über ihr Geschäftsfeld unterhalten.
(2) Vor allem die Kundengruppen des Unternehmens hat er dabei kennen gelernt.
(3) Im Anschluss hat er relevante Daten aus einem SAP-ERP-System und einer Oracle Datenbank geladen.
(4) Mit diesen Daten hat er ein Modell gebaut, welches neue Kunden in die bekannten Kundengruppen einordnet.
(5) Peters Modell wird jetzt im Marketing verwendet, um frühzeitig passende Werbeangebote für neue Kunden zu erstellen.

a) Identifizieren Sie die Phasen eines Data-Science-Projekts in dem obigen Text. Beziehen Sie sich dabei auf die Satznummern. Beschreiben Sie die entsprechende Phase auch kurz allgemein. Sie können sich an Modellen wie CRISP-DM oder KDD orientieren. (8 Punkte)
b) Welche Phase/Phasen fehlt/fehlen in dem Fallbeispiel? Welche negativen Folgen können dadurch entstehen? (4 Punkte)
c) Erklären Sie, welche Art von Modell Peter entwickelt hat (Regression, Clustering oder Klassifikation) und woran dies erkennbar ist. (3 Punkte)

Aufgabe 7: Gemischtes (14 Punkte)

a. Eine Möglichkeit mit fehlenden Werten umzugehen, besteht darin, sie zu entfernen. Welche Gefahr besteht dabei? (2 Punkte) Beschreiben Sie eine Alternative. (2 Punkte)
b. Nennen Sie zwei Beispiele für strukturierte und zwei Beispiele für unstrukturierte Datenquellen. (2 Punkte)
c. Beschreiben Sie zwei Unterschiede zwischen OLAP-Systemen und transaktionsorientierten, operativen Systemen. (4 Punkte)
d. Erklären Sie den Unterschied von Mustererkennung und Prognose. (4 Punkte)

Bonusaufgabe (4 Punkte)

In Ihrem Vortrag Bits and Bias hat Lisa Hanstein als eine mögliche Ursache für einen Algorithmic Bias verzerrte Daten genannt. Beschreiben Sie kurz zwei konkrete Beispiele für diese Form „unfairer IT“.