Otto-von-Guericke Universität Magdeburg
Fakultät für Wirtschaftswissenschaft

Klausur im Fach

Big Data Anwendungen
Wintersemester 2017/2018

Angaben zur Klausur

Prüfer: Dr. Stephan Schosser

Datum: 25. Juli 2017

Prüfungsnummer: 21807

Persönliche Angaben (in Druckbuchstaben ausfüllen)

Nachname: _____ Vorname: _____

Matrikelnummer: _____ Fakultät: _____

Bewertung (wird vom Prüfer ausgefüllt)

Aufgabe	1	2	3	Gesamtpunkte	Note
Punkte					

Zugelassene Hilfsmittel

- Nicht-programmierbarer Taschenrechner
  ohne Kommunikations- oder Datenverarbeitungsfunktion (lt. Aushang des Prüfungsamtes)

Hinweise zur Klausur

- Die Bearbeitungszeit für diese Klausur beträgt 60 Minuten.
- Die Klausur besteht aus 3 Aufgaben, von denen 3 Aufgaben zu bearbeiten sind.
- Die Klausur umfasst 2 Seiten.
- Die Heftung dieser Unterlagen darf nicht gelöst werden.

Hinweise zur Bearbeitung

- Bitte tragen Sie oben auf diesem Deckblatt zuerst Ihre persönlichen Daten ein.
- Bitte prüfen Sie die Vollständigkeit der Klausur.
- Sie sind dafür verantwortlich, dass das Aufsichtspersonal Ihre Klausur erhält.
- Viel Erfolg beim Lösen der Klausuraufgaben!

Aufgabe 1 (Association Rules) **(20 Punkte)**

Gegeben seien folgende Daten:

Brötchen	Brezeln	Brot	Salat	Kuchen	Die Bild
1	1	1	1	0	1
0	1	0	1	1	0
0	0	1	1	0	0
1	0	0	1	0	0
1	1	1	1	1	0
0	0	1	1	1	1

(a) Leiten Sie mit Hilfe des Apriori Algorithmus die Frequent Itemsets ab, die einen Support von mindestens 1/2 besitzen. **(10 Punkte)**

(b) Identifizieren Sie auf der Basis der Frequent Itemsets aus (a) alle Association Rules mit minimaler Confidence von 3/4. **(5 Punkte)**

(c) Diskutieren Sie die Unterschiede der in der Vorlesung besprochenen Algorithmen zur Identifikation von Frequent Itemsets. **(5 Punkte)**

Aufgabe 2 (Clustering) **(20 Punkte)**

Gegeben seien folgende Daten:

Tag	Anzahl Sonnenstunden	Temperatur	Regen
18	10	-1,0	Ja
17	2	5,0	Nein
25	4	1,0	Ja
30	3	2,0	Ja

(a) Nennen Sie alle Attribute für die mit Hilfe des Jaccard-Index keine Abstände ermittelt werden können und begründen Sie kurz. **(2 Punkte)**

(b) Nennen Sie zwei Verfahren zur Transformation von Daten auf ähnliche Wertebereiche. **(2 Punkte)**

(c) Transformieren Sie die Daten so, dass Sie mit Hilfe der Euklidischen Distanz sinnvoll geclustert werden können. **(3 Punkte)**

(d) Nutzen Sie hierarchisch agglomeratives Clustering zum Clustern der Daten aus (c). **(10 Punkte)**

(e) Zeichnen Sie ein Dendrogramm für die Daten aus (d). **(3 Punkte)**

Aufgabe 3 (Klassifikation) **(20 Punkte)**

Gegeben seien folgende Trainingsdaten:

Churn	Anruf im CallCenter	Beschwerde	Zahlungsverzug
Ja	Ja	Ja	Ja
Ja	Nein	Nein	Nein
Ja	Nein	Ja	Ja
Nein	Ja	Nein	Ja
Nein	Nein	Ja	Nein
Nein	Nein	Nein	Ja

(a) Leiten Sie das Attribut Churn mit Hilfe eines Entscheidungsbaums ab. Nutzen Sie hierfür als Splitkriterium den Gini-Index und entwickeln Sie solange neue Knoten bis entweder kein Splitattribut mehr verfügbar ist oder in einem Knoten alle Daten in der gleichen Klasse sind. **(10 Punkte)**

(b) Sagen Sie mit Hilfe des Entscheidungsbaums aus (a) die Empfehlung für folgende Beobachtungen voraus (Ja, Nein, Nein) und (Ja, Ja, Nein). **(6 Punkte)**

(c) Nennen Sie zwei andere Klassifikationsverfahren neben den Entscheidungsbäumen und diskutieren Sie kurz die Vorteile der Verfahren gegenüber Entscheidungsbäumen. **(4 Punkte)**