Otto-von-Guericke Universität Magdeburg
Fakultät für Wirtschaftswissenschaft

Klausur im Fach

Big Data Anwendungen
Sommersemester 2020

Angaben zur Klausur

Prüfer: Dr. Stephan Schosser

Datum: 22. Juli 2020

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 (Clustering) **(20 Punkte)**

Gegeben seien folgende Daten:

Alter	Retouren 2019	Nachfrage 2019
18	50%	550
18	10%	50
24	20%	100
30	0%	250

(a) Nennen Sie zwei Verfahren zur Transformation von Daten auf ähnliche Wertebereiche. Nutzen Sie eines davon und transformieren Sie die Daten so, dass Sie mit Hilfe der euklidischen oder der Manhattan-Distanz sinnvoll geclustert werden können. **(4 Punkte)**

(b) Nutzen Sie kMeans Clustering zum Clustern der Daten aus (a). Nutzen Sie dabei die euklidische Distanz als Distanzmaß und ermitteln sie zwei Cluster. **(10 Punkte)**

(c) Ordnen sie den neuen Datenpunkt: Alter=30, Retouren 2019=30% und Nachfage 2019=100 einem der Cluster aus (b) zu. **(3 Punkte)**

(d) Erläutern Sie kurz den Jaccard-Index und beschreiben Sie, wofür er im Kontext von Clustering genutzt wird. **(3 Punkte)**

Aufgabe 2 (Klassifikation) **(20 Punkte)**

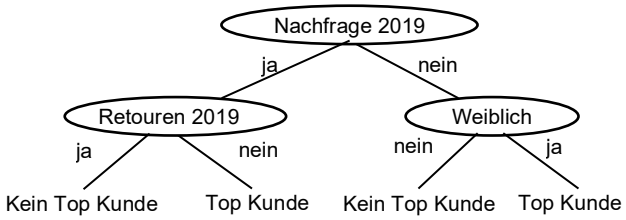
Gegeben seien folgende Test- und Trainingsdaten:

Weiblich	Retouren 2019	Nachfrage 2019	Top Kunde
Ja	Ja	Ja	Nein
Ja	Nein	Nein	Nein
Ja	Ja	Nein	Nein
Ja	Ja	Ja	Nein
Nein	Ja	Ja	Ja
Nein	Nein	Nein	Ja
Nein	Ja	Nein	Ja
Ja	Nein	Nein	Ja

(a) Unterteilen Sie die Daten in einen Test- und einen Trainingsdatensatz mit 25% bzw. 75% der Daten. Gehen Sie davon aus, dass Sie die Eigenschaft „Top Kunde" vorhersagen wollen. Wählen Sie bei der Unterteilung der Daten ein anderes Verfahren als reines zufälliges Ziehen und diskutieren Sie die Vorteile ihres Ansatzes. **(4 Punkte)**

(b) Trainieren Sie auf ihren Trainingsdaten ein Modell zur Vorhersage des Attributs „Top Kunde" mit Hilfe eines Entscheidungsbaums. Nutzen Sie hierfür als Splitkriterium die Entropie und entwickeln Sie nur den Split im Wurzelknoten. **(10 Punkte)**

(c) Gegeben sei folgender trainierter Baum:



Bewerten Sie die Güte des Baums unter Einsatz ihrer Trainingsdaten und zweier Kennzahlen aus der Vorlesung. **(3 Punkte)**

(d) Erläutern Sie kurz, wie die Daten transformiert werden müssen um mit dem kNearest Neighbor verfahren Klassen vorhersagen zu können. **(3 Punkte)**

Aufgabe 3 (Sonstiges) **(20 Punkte)**

(a) Erläutern Sie Link Prediction Algorithmen und deren Nutzen für soziale Netzwerke. **(4 Punkte)**

(b) Erläutern Sie MapReduce und welchen Vorteil dies für Big Data bringt. **(4 Punkte)**

(c) Erläutern Sie die Kölner Phonetik und wie diese helfen kann die Datenqualität zu steigern. **(4 Punkte)**

(d) Erläutern Sie die Unterschiede zwischen Content Based und Collaborativ Filtering. **(4 Punkte)**

(e) Erläutern Sie die Zentralität in sozialen Netzwerken und deren ökonomische Bedeutung. **(4 Punkte)**