Otto-von-Guericke Universität Magdeburg
Fakultät für Wirtschaftswissenschaft

Klausur im Fach

Big Data Anwendungen
Sommersemester 2023

Angaben zur Klausur

Prüfer: Dr. Stephan Schosser

Datum: 29. Juli 2023

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 programmierter Taschenrechner (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 (Klassifikation) **(20 Punkte)**

Gegeben seien folgende Daten über Passagiere der Titanic:

Geschlecht	Gebuchte Klasse	Geschwister	Überlebt
Männlich	1	1	Ja
Weiblich	2	2	Ja
Weiblich	3	4	Nein
Weiblich	3	4	Ja
Männlich	3	2	Nein

(a) Gehen Sie davon aus, dass im Entscheidungsbaum bereits ein Knoten „Geschwister" ermittelt wurde. Dieser besitzt einen linken Ast „<=1" und einen rechten Ast „>1". Ermitteln Sie für den rechten Ast den nächsten Split zur Vorhersage des Attributs „Überlebt". Nutzen Sie hierfür das X^2-Maß. **(10 Punkte)**

(b) Erläutern Sie ob und wenn ja unter welchen Voraussetzungen ein weiterer Knoten „Geschwister" im Entscheidungsbaum auftreten kann. **(3 Punkte)**

(c) Machen Sie für die Beobachtung: Geschlecht – weiblich, gebuchte Klasse – 2, Geschwister – 2 eine Vorhersage. Gehen Sie dabei davon aus, dass die Baumerstellung nach Erstellung des Knotens aus (a) endet. **(2 Punkte)**

(d) Diskutieren Sie wofür ein solches Modell genutzt werden kann. **(3 Punkte)**

(e) Nennen Sie zwei Alternativen zu Entscheidungbäumen zur Klassifikation. **(2 Punkte)**

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

Gegeben seien folgende Daten von Bäumen:

Blütenblattbreite	Blattbreite
2cm	5cm
3cm	6cm
3cm	4cm
1cm	5cm
2cm	1cm

(a) Wenden Sie den k-Means Algorithmus auf die Daten graphisch an. Gehen sie dabei von 2 Clustern aus. **(5 Punkte)**

(b) Führen Sie rechnerisch einen hierarchisch agglomeratives Clustering Verfahren auf die Daten an und nutzen Sie dabei die Manhattan Distanz. **(10 Punkte)**

(c) Übertragen Sie die Daten aus (b) in ein Dendrogramm **(2 Punkte)**

(d) Beschreiben Sie kurz die Vor- und Nachteile von k-Means Algorithmus gegenüber hierarchisch agglomerativem Clustering. **(3 Punkte)**

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

(a) Diskutieren Sie drei Maße, die sie als Alternative zur Akkuratheit einsetzen können und gehen sie dabei darauf ein, warum diese ergänzend sinnvoll sind. **(4 Punkte)**

(b) Diskutieren Sie warum FP Growth ein besserer Algorithmus als der Apriori Algorithmus ist um Frequent Items zu identifizieren. **(4 Punkte)**

(c) Erläutern Sie, wie Collaborative Filtering und Content Based Filtering kombiniert werden können, um dem Cold Start Problem (d.h. fehlenden Daten bei neuen Kunden) zu begegnen. **(4 Punkte)**

(d) Erläutern Sie den Map-Reduce-Ansatz. Gehen Sie dabei besonders darauf ein, wie dieser helfen kann um Maßendaten (Big Data) zu verarbeiten. **(4 Punkte)**

(e) Erläutern Sie, wie sich der Abstand zwischen zwei Beobachtungen mit ausschließlich kategorischen Eigenschaften bestimmen lässt. **(4 Punkte)**