DAV

DEUTSCHE
AKTUARVEREINIGUNG e.V.

Schriftliche Prüfung im Spezialwissen

# Actuarial Data Science Basic

gemäß Prüfungsordnung 4.1
der Deutschen Aktuarvereinigung e. V.

am 27.05.2023

*Hinweise:*

- Als Hilfsmittel ist ein Taschenrechner zugelassen.
- Die Gesamtpunktzahl beträgt 180 Punkte. Die Klausur ist bestanden, wenn mindestens 90 Punkte erreicht werden.
- Bitte prüfen Sie die Ihnen vorliegende Prüfungsklausur auf Vollständigkeit. Die Klausur besteht aus 12 Seiten.
- Alle Antworten sind zu begründen und bei Rechenaufgaben muss der Lösungsweg ersichtlich sein.

*Mitglieder der Prüfungskommission:*

Axel Kiermaier, Dr. René Külheim, Dr. Jonas Offtermatt,
Tobias Renner, Dr. Felix Spangenberg

**Aufgabe 1** [2.1 Datenmanagement 1, 2.3 Informationsverarbeitung in Versicherungen, 3.3 Datenvisualisierung, 4.1 Data Mining 1] (34 Punkte)

Gemeinsam mit Ihren Kollegen möchten Sie die Daten für den Jahresabschluss aufbereiten und analysieren. Sie haben hierfür folgenden Datensatz von Ihren Kolleginnen aus der Leistungsabteilung bekommen:

| Id | Jahr | Monat | Region | Schadenhöhe | Stornoquote |
|---|---|---|---|---|---|
| 1 | 2022 | 7 | Mitte | 10,4 | 4,47 |
| 2 | 2022 | 8 | Mitte | 18,4 | 4,63 |
| 3 | 2022 | 99 | Mitte | 11,6 | - |
| 4 | 2022 | 10 | Mitte | 8,7 | 4,44 |
| 5 | 2022 | 11 | Mitte | 15,5 | 4,74 |
| 6 | 2022 | 12 | Mitte | 10,2 | 4,5 |
| 7 | 22 | 7 | Nird | 1,6 | 2 |
| 8 | 2022 | 8 | Nord | 9 | 3,97 |
| 9 | 2022 | 9 | Nord | 5,6 | 2,3 |
| 10 | 2022 | 10 | Nord | 6,3 | 2,47 |
| 11 | 2022 | 11 | Nord | 10,4 | 5,51 |
| 12 | 2022 | 12 | Nord | 6,8 | 2,64 |
| 13 | 2022 | 7 | Süd | 6,3 | 6,85 |
| 14 | 2022 | 8 | Süd | 12,1 | 3,95 |
| 15 | 2022 | 9 | Süd | NaN | 4.15 |
| 16 | 2022 | 10 | Süd | 12,6 | 3,7 |
| 17 | 2022 | 11 | Süd | 15,6 | 2,2 |
| 18 | 2022 | 12 | Süd | 11 | 4,5 |

a) [9 Punkte] Offensichtlich enthält die Datenlieferung fehlerhafte und fehlende Daten. Nennen Sie drei Methoden, um mit fehlenden Daten umzugehen, und bereinigen Sie die Zeilen mit der Id 3, 7, und 15.

b) [10 Punkte] Visualisieren Sie die Beziehungen zwischen Schadenhöhe, Stornoquote und Region über die Zeit. Verwenden Sie hierfür EINES der auf den der Klausur beigelegten Seiten vorgedruckten Koordinatensysteme.

*Achten Sie darauf, die von Ihnen bearbeitete Seite zusätzlich abzugeben!*

c) [9 Punkte] Nennen und erläutern Sie drei Kriterien für Datenqualität und machen Sie drei konkrete Vorschläge, wie die Datenqualität für den Jahresabschluss erhöht werden kann.

d) [6 Punkte] Um den Jahresabschluss im nächsten Jahr zu beschleunigen, setzen Sie in Ihrem Unternehmen neuerdings auf interaktive Notebooks und

die Methoden des Reproducible Research. Darum sollen die obigen Daten im maschinenlesbaren Format JSON abgespeichert werden. Schreiben Sie die ersten beiden Datenzeilen mit der Id 1 und 2 im JSON-Format auf.

**Lösungsvorschlag:**

a) Mögliche ad-hoc Lösungsansätze (drei hiervon müssen genannt werden):

- Kennzeichnung des Datensatzes als fehlerhaft
- Entfernen des Datensatzes
- Ignorieren/Entfernen des Attributs
- Manuelle Vorgabe fehlender Werte
- Ersatz durch globale Konstante (in R: NA datentypabhängig)
- Ersatz durch Mittelwert / Median
- Ersatz durch Modus (häufigster Wert; bei nicht num. Attributen)
- Ersatz durch Minimum oder Maximum
- Ersatz durch alle möglichen Merkmalsausprägungen (bei diskreten Merkmalen)
- Ersatz durch zufälligen Wert aus anderen Datensätzen
- Ersatz durch interpolierten Wert (z.B. bei Zeitreihen)
- Closest Fit bzw. als Mittelwert unter den ähnlichsten Sätzen („kNN")

Gültig ist auch die Nennung der im Skript nicht genauer erläuterten Methoden MICE = Multiple Imputation by Chained Equations oder EM-Algorithmus = Expectation-Maximization-Algorithmus.

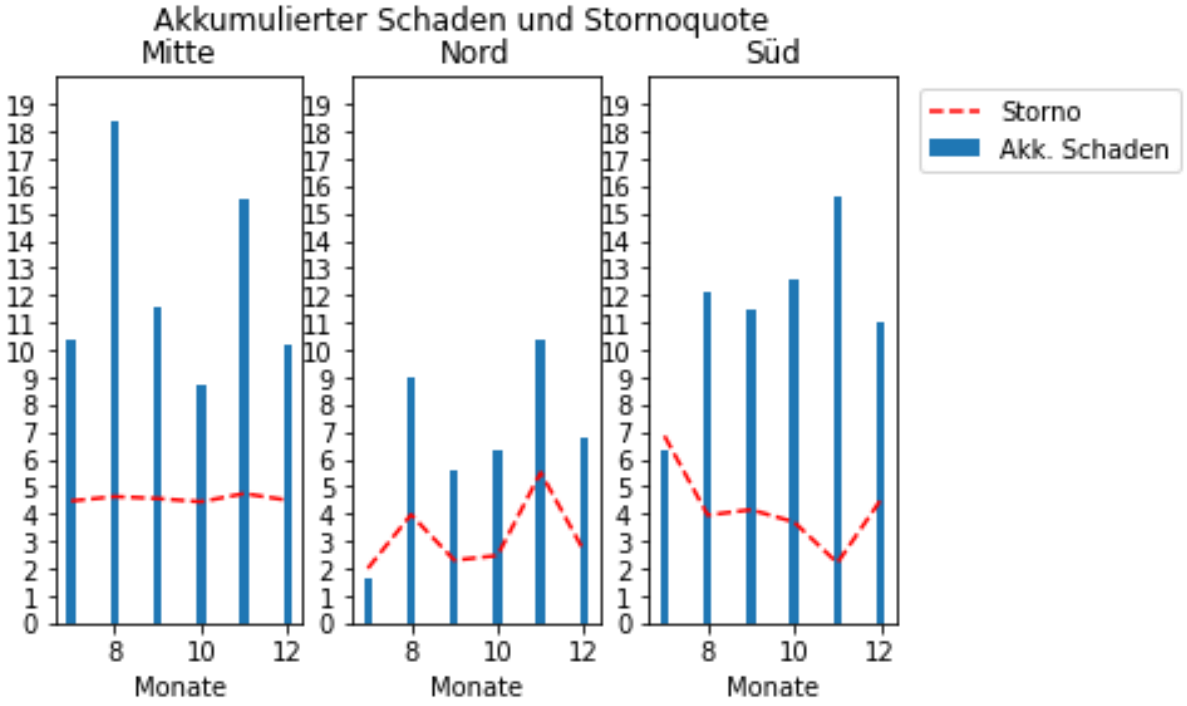
Bereinigung: Offensichtlich handelt es sich bei den Daten zu Id 7 um Tippfehler, welche einfach korrigiert werden können: „Nird" zu „Nord" und „22" zu „2022".

Bei den fehlenden Werten in Id 3 und Id 15 kann bspw. der Mittelwert ergänzt werden, also „-" zu beispielsweise 4,56 (Mittelwert der Monate 7, 8, 10, 11 und 12 der Region Mitte) und „NaN" beispielsweise zu 10,12 (Mittelwert aller Schadenhöhen).

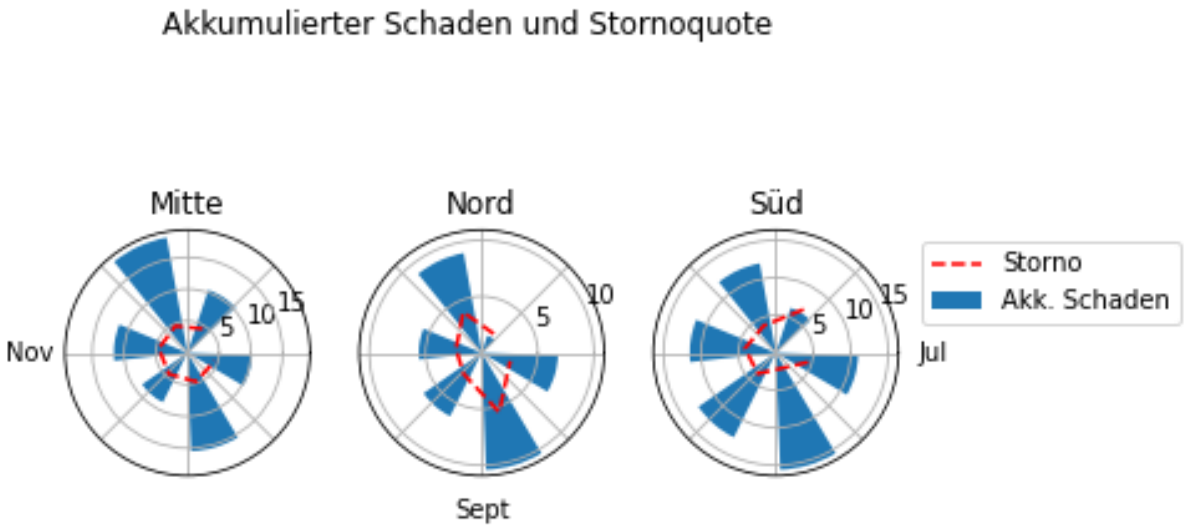
Vorsicht: Bei Id 3 hat sich zusätzlich ein Tippfehler eingeschlichen („99" zu „9"). Bei Id 15 muss man genau hinsehen: Die Stornoquote hat einen Punkt statt einem Komma als Dezimaltrennzeichen.

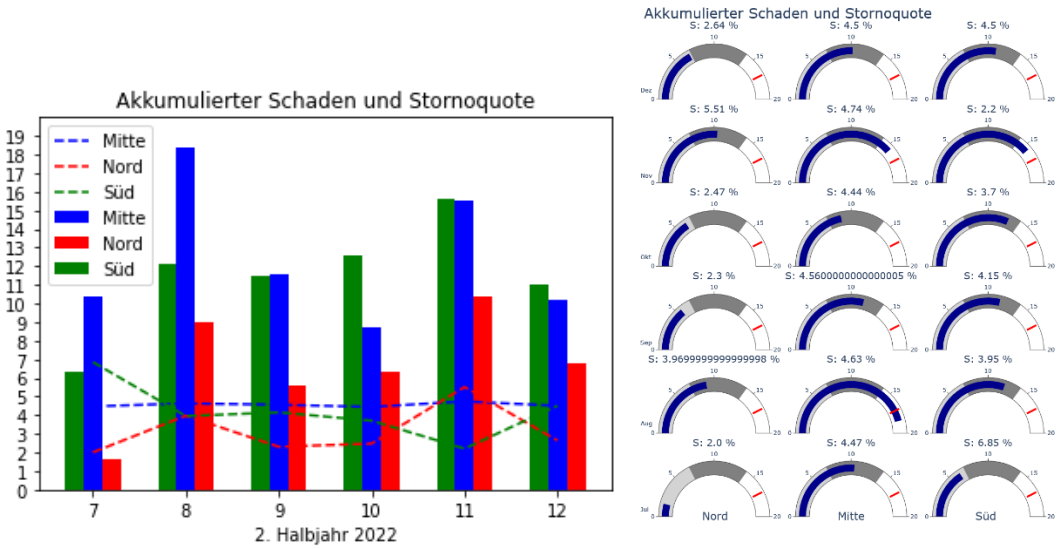
*Zwei Punkte pro Methode und drei Punkte für die korrekte Bereinigung.*

b) Eine mögliche Visualisierung wäre:




Schaubilder weiterer möglicher Visualisierungen:






*Abzüge von Punkten für fehlende Titel, Legenden, Achsenbeschriftungen, etc.*

c) Mögliche korrekte Lösungen sind:

- Aktualität: Alle Datensätze müssen jeweils dem aktuellen Zustand der abgebildeten Realität entsprechen.
- Zuverlässigkeit: Die Entstehung der Daten bzw. die Datenquelle muss nachvollziehbar sein.
- Korrektheit: Die Daten müssen mit der Realität übereinstimmen.
- Konsistenz: Ein Datensatz darf in sich und zu anderen Datensätzen keine Widersprüche aufweisen.
- Vollständigkeit: Ein Datensatz muss alle notwendigen Attribute enthalten.
- Genauigkeit: Die Daten müssen in der jeweils geforderten Exaktheit vorliegen (Beispiel: Nachkommastellen).
- Relevanz: Der Informationsgehalt von Datensätzen muss den jeweiligen Informationsbedarf erfüllen.
- Einheitlichkeit: Die Informationen eines Datensatzes müssen einheitlich strukturiert sein.
- Eindeutigkeit: Jeder Datensatz muss eindeutig interpretierbar sein.

- Verständlichkeit: Die Datensätze müssen in ihrer Begrifflichkeit und Struktur mit den Vorstellungen der Fachbereiche übereinstimmen.
- Redundanzfreiheit: Innerhalb der Datensätze dürfen keine Dubletten vorkommen.

*Drei Kriterien hiervon müssen genannt und kurz erläutert werden.*

*Bei Vorschlägen für die Verbesserung der Datenqualität ist der Kreativität freier Raum gelassen. Mögliche Antworten aus dem Skript wären:*

**Präventive Maßnahmen:**

- Vermeidung der fehlerhaften Erfassung von Daten durch
  - Verwendung von übersichtlichen Auswahllisten
  - Automatische Überprüfung von Eingaben
  - Unwissen zulassen / abgestufte Plausibilisierung
  - Zeitpunkt der Datenerhebung festlegen
- Einführung Datenarchitektur / unternehmensweites Datenmodell
  - Zuordnung der Daten zu Geschäftsdomänen
  - Vermeidung von Redundanz
  - Verwendung von fachlichen Definitionen
  - Semantische Überlagerungen auflösen

**Maßnahmen auf gegebenen Daten:**

- Messen der Datenqualität durch Abgleich gegen Kenngrößen (Ausreißer etc. finden)
- Regelbasiert auf Basis definierter Abweichungstoleranzen

*2 Punkte pro Kriterium und 1 Punkt pro Vorschlag.*

d) Die Speicherung als JSON-String ist nicht eindeutig. Es gibt verschiedenste Möglichkeiten, die zwei Zeilen als JSON zu speichern. Wichtig für eine korrekte Lösung ist, dass die angegebene Kodierung der RFC 7159 der IETF entspricht. Grob zusammengefasst: Nur die Zeichen { }, [ ] , : . und " sind zulässig. Objekte werden in geschweifte Klammern geschrieben, Listen in

eckige Klammern. Eigenschaften von Objekten haben einen Namen und einen Wert, beides getrennt durch einen Doppelpunkt. Zahlen haben Vorkommastellen und Nachkommastellen getrennt durch einen Dezimalpunkt.

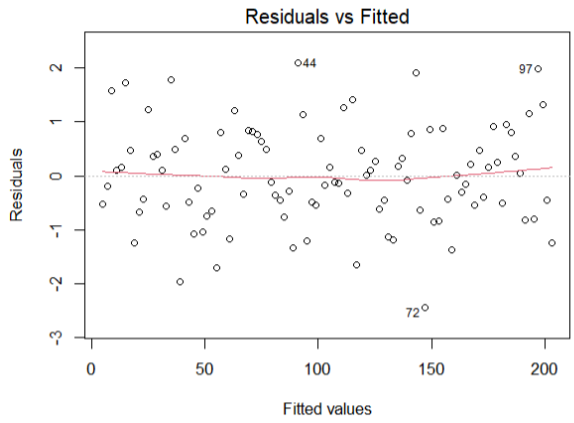
Mögliche korrekte Darstellungen wären somit:

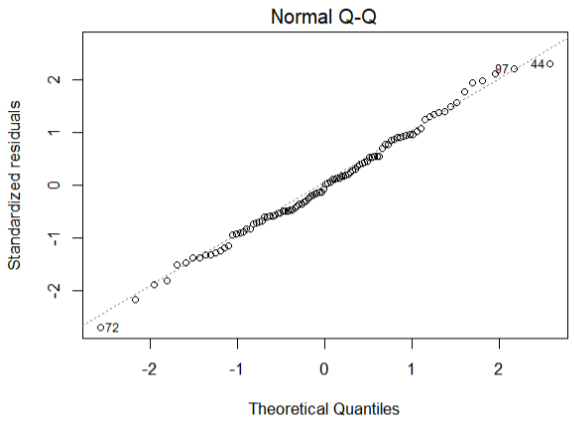
```
Spaltenorientiert:
{"Jahr":{"0":2022,"1":2022},"Monat":{"0":7,"1":8},"Re-
gion":{"0":"Mitte","1":"Mitte"},"Akkumulierter Scha-
den":{"0":10.4,"1":18.4},"Stornoquote":{"0":4.47,"1":4.63}}

Indexorientiert:
{"0":{"Jahr":2022,"Monat":7,"Region":"Mitte","Akkumulierter Scha-
den":10.4,"Stornoquote":4.47},"1":{"Jahr":2022,"Monat":8,"Re-
gion":"Mitte","Akkumulierter Schaden":18.4,"Stornoquote":4.63}}

Zeilenorientiert:
[{"Jahr":2022,"Monat":7,"Region":"Mitte","Akkumulierter Scha-
den":10.4,"Stornoquote":4.47},{"Jahr":2022,"Monat":8,"Re-
gion":"Mitte","Akkumulierter Schaden":18.4,"Stornoquote":4.63}]

Werteorientiert:
[[2022,7,"Mitte",10.4,4.47],[2022,8,"Mitte",18.4,4.63]]

Tabellenorientiert (mit Schema):
{"schema":{"fields":[{"name":"index","type":"inte-
ger"},{"name":"Jahr","type":"integer"},{"name":"Monat","type":"inte-
ger"},{"name":"Region","type":"string"},{"name":"Akkumulierter Scha-
den","type":"number"},{"name":"Stornoquote","type":"number"}],"prima-
ryKey":["index"]},"data":[{"index":0,"Jahr":2022,"Monat":7,"Re-
gion":"Mitte","Akkumulierter Schaden":10.4,"Stornoquote":4.47},{"in-
dex":1,"Jahr":2022,"Monat":8,"Region":"Mitte","Akkumulierter Scha-
den":18.4,"Stornoquote":4.63}]}
```

*Eine korrekte Darstellung ergibt 6 Punkte, für Fehler in der Darstellung werden Punkte abgezogen.*

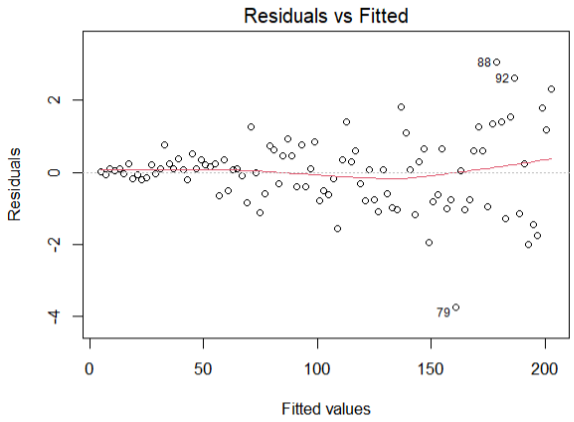
**Aufgabe 2** [3.1 Regressions- und Clustermethoden 1, 1.3 Gesellschaftliches Umfeld und Ethik] (36 Punkte)

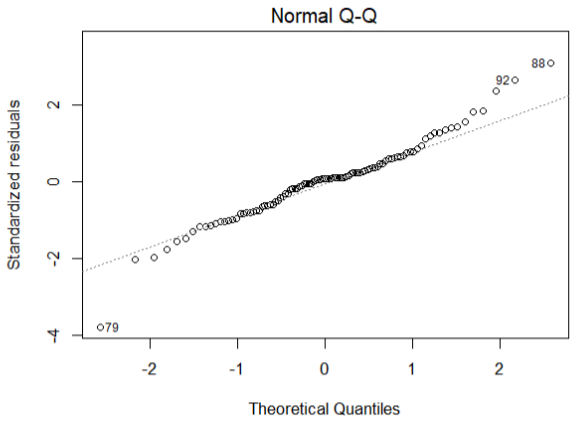
a) [4 Punkte] Sie haben einen Datensatz mit 100 Werten und führen mit R eine lineare Regression unter der Annahme, dass ein linearer Zusammenhang besteht und dass die Störterme unkorreliert und normalverteilt sind, durch. Beurteilen Sie anhand der folgenden Plots, ob die Modellannahmen erfüllt sind. Schlagen Sie, falls aus Ihrer Sicht nötig, einen sinnvollen Ansatz vor, um das Modell zu verbessern.



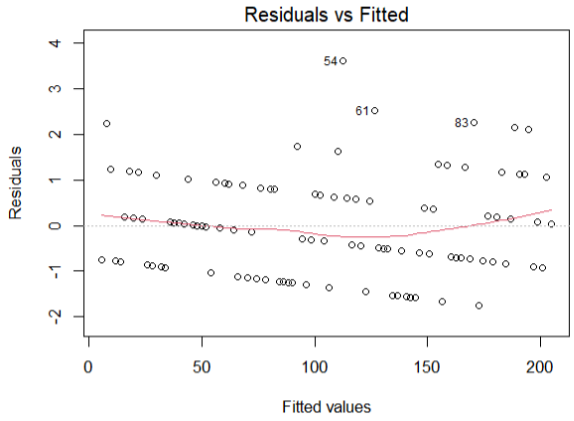



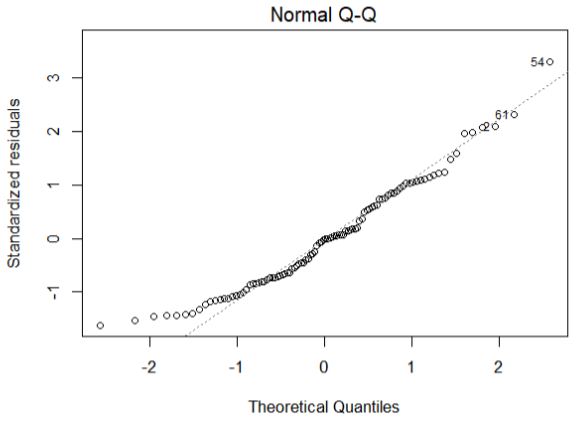

b) [4 Punkte] Sie haben einen Datensatz mit 100 Werten und führen mit R eine lineare Regression unter der Annahme, dass ein linearer Zusammenhang besteht und dass die Störterme unkorreliert und normalverteilt sind, durch. Beurteilen Sie anhand der folgenden Plots, ob die Modellannahmen erfüllt sind. Schlagen Sie, falls aus Ihrer Sicht nötig, einen sinnvollen Ansatz vor, um das Modell zu verbessern.



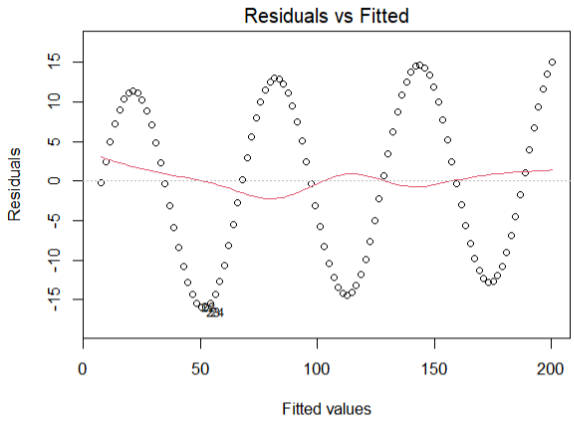



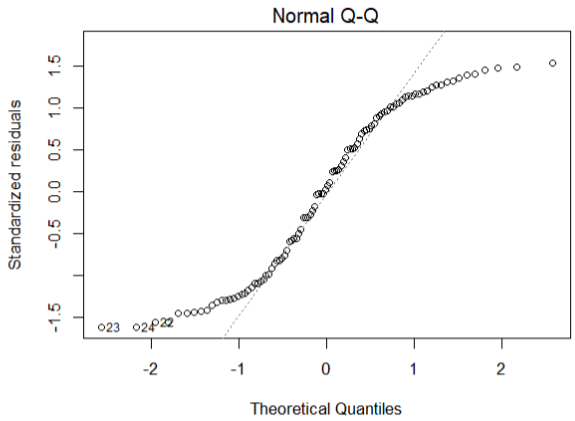
c) [4 Punkte] Sie haben einen Datensatz mit 100 Werten und führen mit R eine lineare Regression unter der Annahme, dass ein linearer Zusammenhang besteht und dass die Störterme unkorreliert und normalverteilt sind, durch. Beurteilen Sie anhand der folgenden Plots, ob die Modellannahmen erfüllt sind. Schlagen Sie, falls aus Ihrer Sicht nötig, einen sinnvollen Ansatz vor, um das Modell zu verbessern.




d) [4 Punkte] Sie haben einen Datensatz mit 100 Werten und führen mit R eine lineare Regression unter der Annahme, dass ein linearer Zusammenhang besteht und dass die Störterme unkorreliert und normalverteilt sind, durch. Beurteilen Sie anhand der folgenden Plots, ob die Modellannahmen erfüllt sind. Schlagen Sie, falls aus Ihrer Sicht nötig, einen sinnvollen Ansatz vor, um das Modell zu verbessern.



e) [20 Punkte] Sie arbeiten bei einem neuen Versicherer, der damit wirbt, dass er besonders faire und verantwortliche Versicherungsprodukte anbietet. Ihr Arbeitgeber möchte, dass Sie vier verschiedene Anwendungen auf ihre Einsetzbarkeit überprüfen. Dabei sollen Sie jeweils eine **kurze** Einschätzung (drei bis fünf Sätze je Anwendung) abgeben, ob der Einsatz dieser Anwendung möglich ist. Gehen Sie hierbei insbesondere auf die folgenden Aspekte ein (je Anwendung 5 Punkte):

- versicherungsmathematische Angemessenheit
- DSGVO-Konformität
- ethische Angemessenheit
- Reputationsrisiko

**Anwendung 1**:

Sie verwenden für die Tarifierung ein GLM.

**Anwendung 2**:

Sie möchten für die Reservierung und das Risikomanagement Machine-Learning-Modelle benutzen, bei denen das Geschlecht berücksichtigt wird.

**Anwendung 3**:

Sie möchten einen Service eines amerikanischen Versicherungsstartups nutzen, das mittels moderner KI-Modelle eine Risikoeinschätzung liefert. Das Machine-Learning-Modell läuft als Blackbox-Modell auf einem amerikanischen Server.

**Anwendung 4**:

Sie sammeln öffentlich zugängliche Informationen über die Versicherungsnehmer unter anderem aus sozialen Medien. Diese Daten sollen dann für den Vertrieb und die Prämienberechnung genutzt werden.

**Lösungsvorschlag:**

a) Der QQ-Plot zeigt, dass die beobachteten Residuen normalverteilt sind. Die Residuen scheinen nicht vom gefitteten Wert abzuhängen. Die Modellannahmen sind erfüllt. Anhand dieser beiden Plots ist keine notwendige Anpassung des Modells erkennbar.

b) Die Varianz der Residuen steigt mit den vorhergesagten Werten. Der QQ-Plot zeigt, dass es Ausreißer gibt. Die Annahme der Normalverteilung ist vermutlich nicht erfüllt. Auch ist die Annahme, dass die Fehlerterme identisch verteilt sind, vermutlich nicht erfüllt. Es bietet sich eine Box-Cox-Transformation an. Alternativ bietet sich an, ein GLM mit einer anderen Verteilungsannahme zu fitten, z.B. Gamma-Verteilung.

c) Anhand des linken Plots ist zu erkennen, dass die Residuen einen linearen Trend annehmen. Die Normalverteilungsannahme ist verletzt und es liegen vermutlich diskrete Beobachtungen vor. Ein Lösungsansatz ist es, ein GLM mit der Poisson-Verteilung zu fitten.

d) Anhand des linken Plots ist zu erkennen, dass ein periodischer Term vorliegt. Ein Erklärungsansatz ist, dass ein periodischer Term im Modell fehlt. Ein anderer Erklärungsansatz ist, dass die Fehlerterme periodisch sind, d.h. die Annahme der Unkorreliertheit verletzt ist. Ein Lösungsansatz ist, das lineare Modell um einen Sinus- und einen Cosinusterm mit entsprechender Frequenz zu ergänzen und das Modell erneut zu fitten.

e) Mögliche Antworten:

**Anwendung 1** darf verwendet werden. Die Verwendung von GLMs ist etablierter aktuarieller Standard. Bei der Modellerstellung muss darauf geachtet werden, dass gesetzliche und ethische Standards wie etwa keine Geschlechterdiskriminierung eingehalten werden. Daten des Versicherungsnehmers dürfen für die Tarifierung benutzt werden, da diese für die Modellerstellung benötigt werden. Es besteht kein Reputationsrisiko.

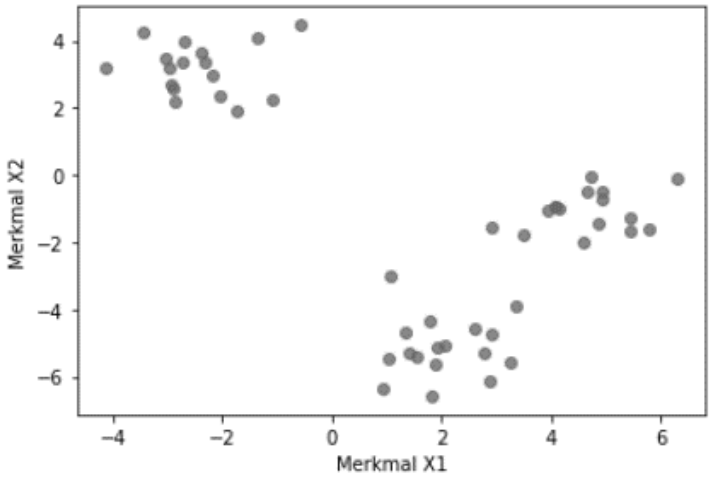
**Anwendung 2** darf verwendet werden. Das Modell ist aktuariell angemessen, da das Geschlecht hier nicht für Prämiengestaltung, Leistungsbearbeitung oder Underwriting genutzt wird. Die Verwendung der Daten für Reservierung und Risikomanagement ist zulässig, da der Versicherer eine rechtliche Verpflichtung dazu hat. Es muss sichergestellt werden, dass die Verwendung des Geschlechts nicht zu indirekter Benachteiligung führt, etwa über eine höhere Überschussbeteiligung in der Lebensversicherung. Es besteht ein kleines Reputationsrisiko, falls die Öffentlichkeit erfährt, dass das Geschlecht in Machine-Learning-Modellen benutzt wird.

**Anwendung 3** darf nicht verwendet werden. Die Übermittlung von Daten in die USA ist aus DSGVO-Sicht hoch problematisch. Außerdem wird hier ein externes Modell benutzt. Dies ist aus aktuarieller und ethischer Sicht problematisch, da das Modell nicht verstanden wird. Aus diesen beiden Gründen besteht auch ein hohes Reputationsrisiko.

**Anwendung 4** darf nicht verwendet werden. Die gesammelten Daten werden ohne Einwilligung erhoben und dürfen daher nicht genutzt werden. Die unautorisierte Datensammlung ist auch aus ethischer Sicht problematisch und birgt ein hohes Reputationsrisiko. Außerdem werden hier Daten erhoben, die keine versicherungsmathematische Relevanz haben oder aufgrund von Diskriminierungsverboten nicht genutzt werden dürfen.

**Aufgabe 3** [3.1 Regressions- und Clustermethoden 1] (40 Punkte)

Im Rahmen einer Bestandsanalyse sollen Sie eine Clusterung durchführen. Hierzu soll der k-Means-Algorithmus verwendet werden. Grundlage der Clusterung sind Daten mit den zwei Merkmalen X1 und X2, wie in der folgenden Grafik dargestellt:




In der folgenden Tabelle sind fünf exemplarische Datensätze aufgelistet:

| ID | X1 | X2 |
|---|---|---|
| 1 | 5 | 0 |
| 2 | -0.58 | 4.46 |
| 3 | -2.74 | 3.34 |
| 4 | 4.08 | -0.94 |
| 5 | -4.13 | 3.21 |

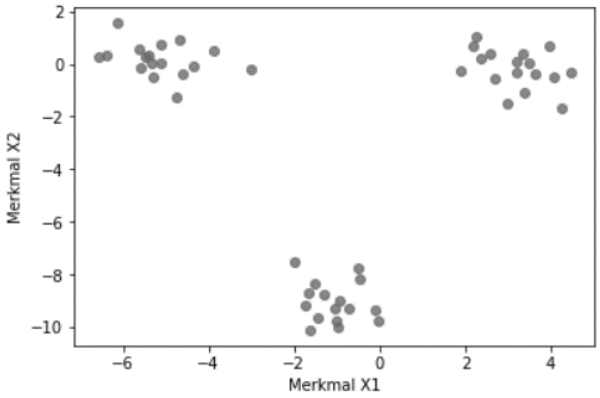
a) [10 Punkte] Beschreiben Sie kurz den k-Means-Algorithmus. Erläutern Sie hierbei die verschiedenen Schritte des Algorithmus, wobei Sie eine geeignete Annahme über die Position der Centroiden der ausgewählten Cluster treffen. Berechnen Sie für den Datensatz mit der ID1 die erste Iteration der k-Means-Clusterung - soweit dies mit einem einzelnen Datensatz möglich ist.

Durch Anpassungen der Datengrundlage ergeben sich jeweils vier neue Datensätze als Grundlage des k-Means-Algorithmus. Hierbei erfolgt die Anpassung jeweils auf Grundlage der Daten aus der Aufgabenstellung (vor der Aufgabe a).

i. Hinzunahme der drei Datensätze:

| ID | X1 | X2 |
|---|---|---|
| 11 | 5.187 | -0.124 |
| 12 | 12.974 | 14.643 |
| 13 | -1.514 | 4.947 |

ii. Hinzunahme der folgenden Datensätze:




iii. Hinzunahme der Datensätze

| ID | X1 | X2 |
|---|---|---|
| 11 | -1.09 | 2.25 |
| 12 | -0.58 | 4.46 |
| 13 | -2.74 | 3.34 |

iv. Hinzunahme von neuen Merkmalen X3 bis X20

b) [12 Punkte] Benennen und erläutern Sie für jede der Datenanpassungen (i. bis iv.), ob und in welcher Form die Anpassung der Datengrundlage Auswirkung auf die Ausführung des k-Means-Algorithmus hat bzw. haben könnte. Beschreiben Sie hierzu die Auffälligkeiten in den Daten und erläutern Sie sinnvolle und notwendige Anpassungen bei der Ausführung oder im Vorfeld der Ausführung des k-Means-Algorithmus.

c) [6 Punkte] Nennen und erläutern Sie kurz drei grundverschiedene Anwendungsbereiche für Clusterverfahren im Versicherungsbereich.

d) [12 Punkte] Eine neben k-Means und seinen Varianten sehr verbreitete Klasse von Clusterverfahren sind hierarchische Clusterverfahren.

Welche zwei grundlegenden Ansätze gibt es bei hierarchischen Clusterverfahren? Erläutern Sie die Grundidee dieser beiden Ansätze in wenigen Sätzen und beschreiben Sie dann einen der beiden Ansätze in algorithmischer Form. Erläutern Sie in diesem Zusammenhang die sog. Complete- und Single-Strategie.

Wie lassen sich die Ergebnisse einer hierarchischen Clusterung visualisieren und inwiefern ergibt sich hieraus ein wichtiger Vorteil gegenüber k-Means?

**Lösungsvorschlag:**

a) Der k-Means-Algorithmus ist ein mathematisches Verfahren zur Gruppierung (Clusterung) von Objekten in einem n-dimensionalen Raum und gehört zu den Verfahren des Unsupervised Learnings. Ziel der Clusterung ist es, Datenpunkte in k homogene Gruppen einzuteilen.

Der k-Means-Algorithmus erfolgt in folgenden Schritten:

1. Festlegung der Anzahl k der Cluster.
2. Festlegung der initialen Lage des Clusterzentrums (centroid) für jedes Cluster.
3. Zuordnung jedes Datenpunkts zu dem nächstliegenden Clusterzentrum unter Verwendung eines Abstandsmaßes (z.B. der euklidischen Distanz).
4. Neuberechnung der Clusterzentren (centroid) auf Basis der zugeordneten Datenpunkte.
5. Wiederholung Schritt 3 und 4 bis sich die Lage der Clusterzentren nicht mehr ändert oder ein definiertes Abbruchkriterium erreicht ist.

Im Folgenden ist die erste Iteration der k-Means-Clusterung für die gegebenen Datensätze gerechnet:

1. Festlegung k = 3
2. Initiale Festlegung der Clusterzentren:
   - [-2, 3]
   - [2, -5]
   - [5, 0]
3. Die Zuordnung der Datenpunkte zu den Clusterzentren soll über die euklidische Distanz erfolgen. Für die Datenpunkte p und d berechnet sich die euklidische Distanz wie folgt:

$$d(p,d) = \sqrt{\sum_{i=1}^{n}(p_i - d_i)^2}$$

Für den ersten Datensatz (ID1 = [5, 0]) ergeben sich die folgenden Distanzen zu den Clusterzentren:

- $Distanz\ zu\ [-2{,}3] = \sqrt{(-2-5)^2 + (3-0)^2} = \sqrt{(-7)^2 + (3)^2} = \sqrt{49+9} = \sqrt{58} = 7{,}62$

- $Distanz\ zu\ [2,-5] = \sqrt{(2-5)^2+(-5-0)^2} = \sqrt{(-3)^2+(-5)^2} = \sqrt{9+25} = \sqrt{34} = 5{,}83$
- $Distanz\ zu\ [5{,}0] = \sqrt{(5-5)^2+(0-0)^2} = \sqrt{(0)^2+(0)^2} = \sqrt{0+0} = \sqrt{0} = 0$

Der Datenpunkt [5,0] ist damit dem am nächsten liegenden Clusterzentrum [5,0] zuzuordnen.

4. Für die Anpassung der Clusterzentren ist die Ermittlung der Distanz für alle Datensätze notwendig. (Auch für die Anpassung eines einzigen Clusterzentrums benötigt man die Distanzen sämtlicher diesem Cluster zugeordneten Datensätze.) Aus diesem Grund ist eine Anpassung der Clusterzentren auf Basis der Werte eines einzelnen Datensatzes nicht möglich.

Eine andere Wahl von k und der Clusterzentren ist möglich.

b) Für die Anpassungen i. bis iv. gilt:

i. Der zweite Datenpunkt ID2 ist ein Ausreißer und eine eindeutige Zuordnung des Datenpunktes zu einem Cluster ist nicht möglich. Durch den Ausreißer besteht die Gefahr, dass ein Clusterzentrum verzerrt wird oder dass durch den Ausreißer ein eigenes Cluster erzeugt wird. Vor der Anwendung des k-Means-Algorithmus sollte eine Datenbereinigung durchgeführt werden und der Ausreißer sollte entfernt werden. Je nach konkreter Anwendung kann der Ausreißer ggf. nach dem Clustern wieder hinzugefügt werden.

ii. Durch die Hinzunahme der Datenpunkte erhöht sich die Anzahl der Cluster von 3 auf 6 Cluster. Ohne eine Anpassung der Anzahl der Cluster liefert der k-Means eine unzureichende Clusterung und innerhalb der Cluster entsteht eine hohe Streuung. Für die Anwendung der Clusterung sollte daher die Anzahl der Cluster k auf 6 gesetzt werden.

iii. Die Hinzunahme der Datenpunkte hat keine signifikante Auswirkung auf die k-Means-Clusterung. Eine Anpassung des k-Means-Algorithmus oder eine andere Anpassung vor Anwendung des Algorithmus ist nicht erforderlich.

iv. Durch die Hinzunahme der Merkmale besteht die Gefahr, dass sich die Datenpunkte verstärkt / zu stark im n-dimensionalen Raum verteilen. Hierdurch besteht die Gefahr bei distanzbasierten Verfahren wie dem k-Means-Algorithmus, dass dieser nicht konvergiert und keine Clusterung durchgeführt werden kann. Vor der Anpassung des k-Means-Algorithmus

sollte nach Möglichkeit eine Dimensionsreduzierung durchgeführt werden. Dies ist aber abhängig vom jeweiligen Anwendungsfall und der erwarteten/angestrebten Clusteranzahl k.

c) *Hier einige Beispiele für den Einsatz von Clusterverfahren im Versicherungsbereich. Für die volle Punktzahl sind drei möglichst verschiedene Anwendungsbereiche zu nennen.*

- Marktanalyse/Marktsegmentierung:
  - Bestimmung von Zielgruppen für neue Produkte
  - Identifizierung von Gruppen in sozialen Netzwerken
- Ermittlung von Risikofaktoren:
  - Ermittlung von Typ- und Regionalklasse für die Kfz-/Sachtarifierung
  - Ermittlung von Genen/Genkombinationen, die für eine bestimmte Krankheit (z.B. Krebs) verantwortlich sind
- Risikomanagement / Projektionsrechnungen:
  - Ermittlung von Stresstest-Szenarien
  - Bestandsverdichtung insb. in der Lebensversicherung

d) Grundsätzlich sind bei hierarchischen Cluster-Methoden zwei **Ansätze** zu unterscheiden:

- Bei agglomerativen Verfahren bildet zunächst jede einzelne Beobachtung ein eigenes Cluster. Diese Cluster werden dann schrittweise abstandsbasiert zu größeren Clustern zusammengefasst.
- Divisive Clusterverfahren hingegen gehen so vor, dass alle Datenpunkte zunächst ein einziges großes Cluster bilden, welches dann schrittweise in kleinere Unter-Cluster zerlegt wird.

Der allgemeine agglomerative **Algorithmus** sieht so aus:

1. Vorgabe von n Beobachtungen und einem Abstands-/Ähnlichkeitsmaß.

2. Bildung von n Clustern bestehend jeweils aus genau einer Beobachtung.

3. Für i = n, n - 1, …, 2:

   Berechne die paarweisen Abstände/Ähnlichkeiten der i Cluster unter Verwendung des gewählten Abstands-/Ähnlichkeitsmaßes.

   Identifiziere die beiden ähnlichsten Cluster und fasse diese zu einem Cluster zusammen.

**Zur Complete- und Single-Strategie:**

Hierbei handelt es sich um zwei mögliche Strategien (es gibt noch einige weitere), um die Abstände von zwei Clustern im Rahmen des Algorithmus zu bestimmen.

– Complete: Berechnet paarweise Abweichungen zwischen allen Beobachtungen (Datenpunkten) der beiden Cluster und wählt den größten Abstand.

– Single: Berechnet paarweise Abweichungen zwischen allen Beobachtungen (Datenpunkten) der beiden Cluster und wählt den kleinsten Abstand.

**Visualisierung:**

Die Ergebnisse eines hierarchischen Clusterverfahrens können gut anhand einer Baumstruktur visuell dargestellt werden. Die Darstellung eines solchen Baumes wird Dendrogramm genannt.

Die Anzahl k der Cluster muss im **Unterschied zu k-Means** bei hierarchischen Verfahren nicht vor dem Clustering vorgegeben werden. Stattdessen kann nach dem Clustering durch geeignete Wahl der Baumtiefe indirekt die Anzahl der Cluster bestimmt werden. Die Entscheidung für ein anderes k macht im Unterschied zu k-Means keine erneute Durchführung des Algorithmus erforderlich.

**Aufgabe 4** [3.1 Regressions- und Clustermethoden 1, 1.3 Gesellschaftliches Umfeld und Ethik, 4.2 Analytics 1 (4.2)] (35 Punkte)

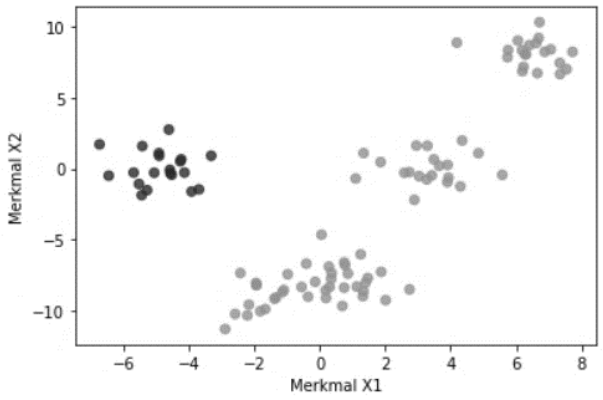
a) [7 Punkte] Definieren Sie zunächst die beiden Begriffe Klassifikation und Clusteranalyse kurz und prägnant. Erläutern Sie dann die wichtigste Gemeinsamkeit der beiden Verfahren sowie die grundlegenden Unterschiede. Gehen Sie dabei insbesondere auf den Unterschied zwischen Klassen (im Sinne der Klassifikation) und Cluster (im Sinne der Clusteranalyse) ein.

b) [14 Punkte] Ihr Vorstand erwägt, zur Vereinfachung der Risikoprüfung für klassische Lebensversicherungen eine Software anzuschaffen, die auf Ihnen nicht im Detail bekannten Klassifikationsalgorithmen basiert.

Erläutern Sie, welche möglichen Auswirkungen diese Anwendung von Machine Learning für Ihr Unternehmen, für Ihren Vertragsbestand und für die Gesellschaft haben kann. Nennen Sie hierzu für jede dieser drei Kategorien mindestens zwei Auswirkungen, möglichst je eine positive und negative; machen Sie dabei klar, was Sie als positiv oder als negativ einstufen.

Welche Argumente erachten Sie für besonders schwerwiegend und welche Empfehlung sprechen Sie deshalb aus?

c) [14 Punkte] In einer Bestandsanalyse steht Ihnen der nachfolgende Datensatz mit den beschreibenden Merkmalen X1 und X2 zur Verfügung:




Hierbei gilt ● Y=0 ● Y=1 . Y beschreibt hierbei den Eintritt eines Ereignisses.

Durch die Erstellung eines Modells mit einer logistischen Regression wurden die folgenden Parameter ermittelt:

Intercept: $ß_0 = 2{,}169$

Koeffizienten: $ß_1 = 1{,}466$, $ß_2 = -0{,}059$

Zur Überprüfung des Modells steht Ihnen der nachfolgende exemplarische Datensatz zur Verfügung. Hierbei wurden für die Datensätze ID11 bis ID18 bereits Prognosewerte für Y ermittelt.

| ID | X1 | X2 | Y Ist | Y Prognose |
|---|---|---|---|---|
| 11 | -5 | -10 | 0 | 0 |
| 12 | -4 | 2 | 1 | 0 |
| 13 | -3 | 0 | 0 | 0 |
| 14 | -2 | 5 | 0 | 0 |
| 15 | -1 | 5 | 1 | 1 |
| 16 | 0 | 3 | 1 | 1 |
| 17 | 1 | 5 | 0 | 1 |
| 18 | 2 | 10 | 1 | 1 |
| 19 | 0 | 0 | 0 | |
| 20 | 1 | 1 | 1 | |

Erstellen Sie für den gegebenen Testdatensatz eine Confusion Matrix und bewerten Sie die Prognosegüte. Führen Sie hierzu die Prognose mit der logistischen Regression für die Datensätze mit der ID19 und ID20 durch und bewerten Sie die Prognosegüte mit drei verschiedenen Kennzahlen. Erläutern Sie hierbei die Bedeutung der Kennzahlen.

Hinweis: Es gilt *exp(-2,17) = 0,11; exp(-3,58)=0,03.*

**Lösungsvorschlag:**

a) Klassifikation und Clusteranalyse sind Verfahren des maschinellen Lernens.

**Klassifikation**: Hier geht es um das Lernen einer Menge von Regeln, die Objekte aufgrund ihrer Attribute/Merkmale vorgegebenen Klassen zuordnen.

**Clusteranalyse**: Hier ist das Ziel die Zusammenfassung ähnlicher Objekte in vorab nicht bekannte Teilmengen (Cluster).

**Gemeinsamkeit**: Es geht in beiden Fällen darum, den gesamten Beobachtungsraum (möglichst) vollständig in Klassen bzw. Cluster zu zerlegen. Dabei sind die in einer Clusteranalyse gebildeten Cluster stets disjunkt, die Klassen einer Klassifikation häufig disjunkt (Anmerkung: Sind nicht alle Klassen disjunkt, spricht man von multi-label classification).

**Unterschiede**:

| | Klassifikation | Clusteranalyse |
|---|---|---|
| Lernverfahren | überwacht | unüberwacht |
| Analyseart | prädiktiv | deskriptiv |
| Namensgebung der Teilmengen | Klassen haben (vorgegebene) Namen | Cluster werden gefunden und haben keine Namen |

b) Mögliche Auswirkungen:

*Die folgende Darstellung enthält zur Illustration mehr als eine (mögliche) Antwort in den sechs Fällen. Für das Erreichen der vollen Punktzahl sind insgesamt sechs verschiedene Antworten nötig (für jede der drei Kategorien mindestens zwei Auswirkungen, möglichst je eine positive und negative).*

- Auswirkungen auf das Unternehmen:

- Positiv:

  Reputationsgewinn (das Unternehmen gilt als „modern" und „innovativ").

  Niedrigere Kosten und höhere Gewinne durch Einsparungen in den Prozessen (keine/weniger Risikoprüfer nötig).

  Falls der Ansatz tatsächlich zum angestrebten Effizienzgewinn führt, kann sich hieraus ein Wettbewerbsvorteil für den frühzeitigen Anwender einer neuen Technologie ergeben.

- Negativ:

  Reputationsverlust (das Unternehmen gilt als „Datenkrake").

  Hohe Initialkosten / Erstaufwand in der IT für die Umsetzung / Know-how-Aufbau.

  Auf Dauer evtl. niedrigere Prämieneinnahmen, weil bevorzugt die Kunden mit günstigeren Prämien kommen.

  Aufwand, um dem Kunden im Zweifelsfall die Klassifikationsergebnisse der Fremdsoftware zu erklären. Wie kann dieses neue Konzept dem Kunden vermittelt werden, um Vertrauen zu schaffen? Geeignetes Marketing zu entwickeln.

- Auswirkungen auf den Vertragsbestand:

  - Positiv:

    Selektion guter Risiken.

  - Negativ:

    Ggf. werden nur noch gute Risiken angenommen. Hierdurch ist eine Bewertung des Bestands anhand der Erfahrungswerte der Vergangenheit zumindest für eine Übergangszeit nicht in gewohnter Qualität möglich.

    Die Bestandszusammensetzung kann sich je nach angewandtem ML-Algorithmus verändern. Daher ist darauf zu achten, dass diese Veränderung nicht in eine ungewünschte Richtung geht.

- Auswirkungen auf die Gesellschaft:
    - Positiv:

      … (es fällt schwer, positive Auswirkungen auf die Gesellschaft zu finden; *daher sind hier ggf. zwei negative Auswirkungen zu benennen*).

    - Negativ:

      Teile der Bevölkerung sind nicht mehr versicherbar / Personen-Gruppen könnten durch solche Algorithmen benachteiligt werden (Diskriminierung).

      Gesellschaftliche Spaltung in versicherbare und nicht versicherbare Personen.

      Es wird noch schwerer, dem Kunden das Produkt Lebensversicherung zu erklären und damit sinkt auch die Akzeptanz des Produktes LV.

      Datenschutz? / Beziehung der Gesellschaft zum Datenschutz würde sich verändern.

      Vorsorgeuntersuchungen werden nicht mehr durchgeführt, aus Angst vor Ausschluss.

Als besonders problematisch ist die Gefahr der Diskriminierung von Personengruppen einzustufen. Diese scheinbar rein ethische Problematik kann sich leicht zu einem gravierenden Reputationsrisiko auch für das Unternehmen entwickeln. Möglicherweise öffnet sich auch die gesamte LV-Branche für eine ML-basierte Risikoprüfung. Dies könnte sich dann aber auch zu einem Reputationsproblem der gesamten Branche auswachsen. Daher ist aus ethischen und wirtschaftlichen Gründen von der geplanten Softwareeinführung abzuraten.

*Anm.: Dies ist nur ein Beispiel für eine mögliche Empfehlung. Anders gelagerte Empfehlungen sind bei entsprechender Begründung ebenfalls möglich.*

c) Zur Ermittlung der Wahrscheinlichkeit für Y=1 mittels einer logistischen Regression gilt:

$$P(Y=1) = \frac{1}{1+\exp(-y)} = \frac{1}{1+\exp(-(ß_0 + ß_1 x_1 + ß_2 x_2))}$$

Für den Datensatz ID19 gilt:

$$P(Y=1) = \frac{1}{1+\exp\left(-(2{,}169+1{,}466*0-0{,}059*0)\right)} = \frac{1}{1+\exp\left(-2{,}169\right)} = \frac{1}{1{,}11} = 0{,}897$$

Für den Datensatz ID20 gilt:

$$P(Y=1) = \frac{1}{1+\exp\left(-(2{,}169+1{,}466*1-0{,}059*1)\right)} = \frac{1}{1+\exp\left(-3{,}576\right)} = \frac{1}{1{,}03} = 0{,}973$$

Mit der Wahl einer sinnvollen Wahrscheinlichkeitsstufe von 0,5 = 50 % (=Cutoff) wird für die Datensätze ID 19 und ID 20 Y=1 prognostiziert.

Mit den tatsächlichen und prognostizierten Werten für die Datensätze ID 11 bis ID 20 ergibt sich die folgende Confusion Matrix:

| | | Vorhersage | |
|---|---|---|---|
| | | Y=1 | Y=0 |
| Aktueller Wert | Y=1 | 4 | 1 |
| | Y=0 | 2 | 3 |

Hierbei gilt:

- TP (True Positive) = 4
- TN (True Negative) = 3
- FP (False Positive) = 2
- FN (False Negative) = 1

Die Bewertung der Prognosegüte soll über die Kennzahlen Sensitivität (Recall / True Positive Rate), Spezifität (True Negative Rate) und Accuracy erfolgen:

- $\text{Sensitivität} = \frac{TP}{TP+FN} = \frac{4}{5} = 80\%$
- $\text{Spezifität} = \frac{TN}{TN+FP} = \frac{3}{5} = 60\%$
- $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{7}{10} = 70\%$

Die Kennzahl „Sensitivität" beschreibt hierbei den Anteil der korrekt vorhergesagten Datensätze, bei denen das Ereignis eingetreten ist (d.h. $Y_{Ist} = 1$).

Die Kennzahl „Spezifität" beschreibt hierbei den Anteil der korrekt vorhergesagten Datensätze, bei denen das Ereignis nicht eingetreten ist (d.h. $Y_{Ist} = 0$).

Die Kennzahl „Accuracy" beschreibt hierbei den Anteil der insgesamt korrekt vorhergesagten Datensätze (für $Y_{Ist} = 0$ oder $Y_{Ist} = 1$).

*Alternativ ist auch die Verwendung anderer Kennzahlen möglich.*

**Aufgabe 5** [3.1 Regressions- und Clustermethoden 1, 2.2 Datenverarbeitungstechnologien 1] (35 Punkte)

a) [5 Punkte] Folgender vereinfachter Datensatz liegt Ihnen vor:

| **Response** | **Feature 1** | **Feature 2** |
|---|---|---|
| Ja | Ja | Nein |
| Ja | Nein | Ja |
| Nein | Ja | Nein |
| Nein | Ja | Ja |
| Ja | Ja | Ja |
| Nein | Nein | Nein |
| Nein | Ja | Nein |

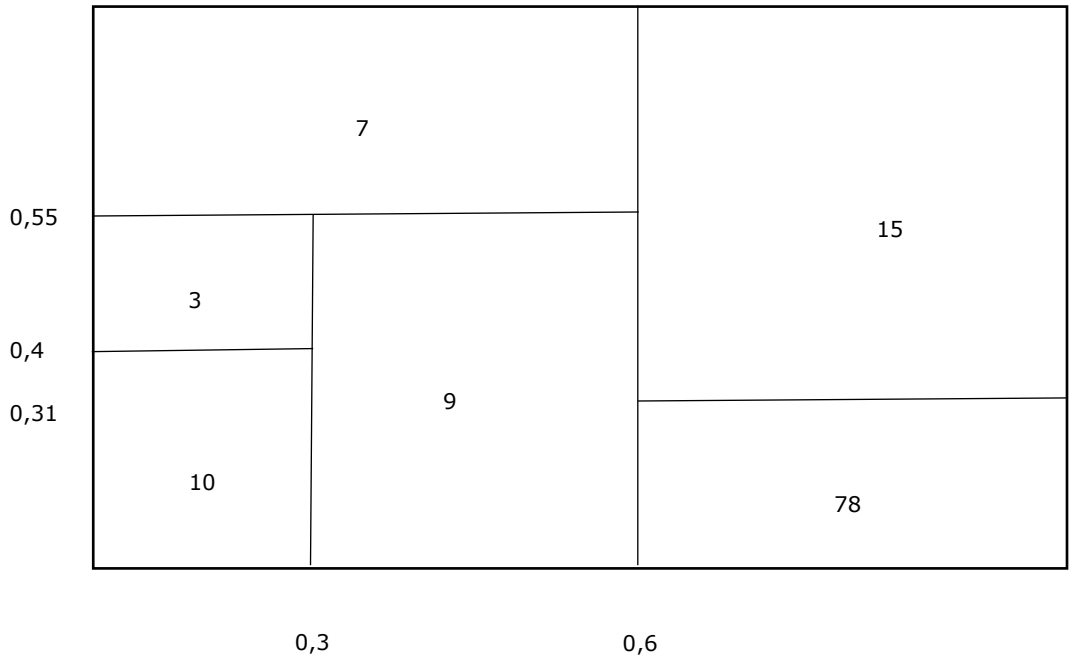
Zur Modellierung des Datensatzes wollen Sie einen einfachen Entscheidungsbaum benutzen. Leiten Sie zunächst aus der allgemeinen Formel für den Gini-Index,

$$G = \sum_{k=1}^{K} \hat{p}_k(1 - \hat{p}_k)$$

eine Berechnungsformel für den Fall von K=2 Klassen her. Im Fall von zwei Klassen bezeichne $\hat{p}_1 = p$ die Wahrscheinlichkeit für „Ja" und $\hat{p}_2 = q = 1 - p$ die Wahrscheinlichkeit für „Nein". Vereinfachen Sie die Formel so weit wie möglich.

b) [15 Punkte] Für den Aufbau eines (einfachen) Entscheidungsbaumes (für den Datensatz aus Teil a) ) ist zu entscheiden, ob der erste Split mit Feature A oder Feature B erfolgen soll. Dazu erzeugen Sie zwei separate Bäume, einmal mit Feature A und einmal mit Feature B an der Spitze des Baumes („Root Node"). Ordnen Sie die Response je nach Wert des Features dem entsprechenden Blatt zu. Berechnen Sie für beide Blätter den Gini-Index mit der Formel aus Teil a). Berechnen Sie im Anschluss den Gini-Index für die beiden Bäume als gewichtetes Mittel. Welches Feature sollte damit im „Root-Node" verortet werden? (Hinweis zum gewichteten Mittel: Mit $n$ Anzahl Personen Blatt 1, $m$ Anzahl Personen Blatt 2, $G_1$ bzw. $G_2$ der jeweilige Gini-Index. Dann lautet die Formel: $\frac{n}{n+m} G_1 + \frac{m}{n+m} G_2$).

c) [5 Punkte] Die folgende Abbildung zeigt eine Partitionierung des Beobachtungsraumes anhand von zwei Features (horizontal: X, vertikal: Y). Rekonstruieren Sie daraus den zugehörigen Entscheidungsbaum:




Hinweis: Achten Sie darauf, dass die jeweiligen Werte für die Endknoten (wie z.B. 7, 15 oder 78) nur einmal im Baum vorkommen.

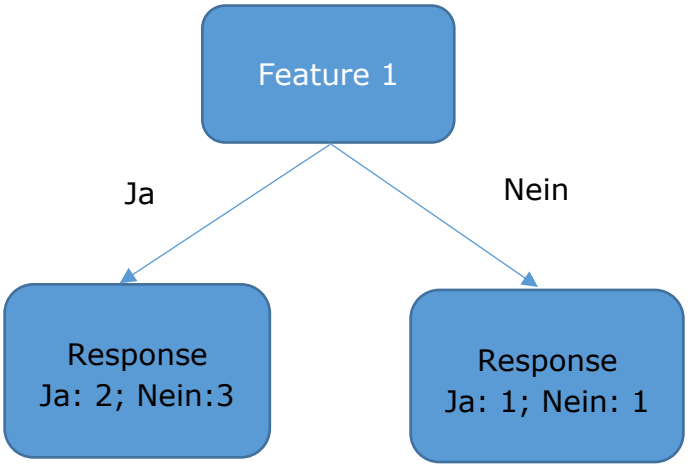
d) [10 Punkte] Für eine geeignete Implementierung einer Data Science-Anwendung machen Sie sich Gedanken zur Skalierbarkeit und zum Testen. Erläutern Sie die Begriffe horizontale und vertikale Skalierbarkeit und nennen Sie Beispiele hierzu. Erläutern Sie im Anschluss die prinzipielle Vorgehensweise beim Test Driven Development (TDD), indem Sie auf den entsprechenden Prozess bei der Softwareentwicklung eingehen.

**Lösungsvorschlag:**

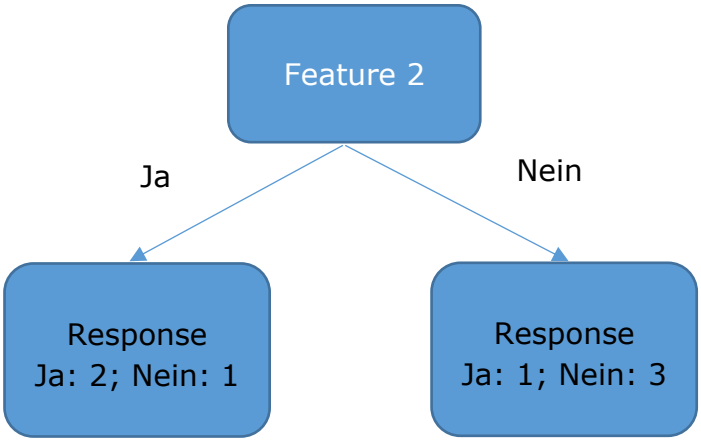
a) Anhand der Response ist erkennbar, dass es sich hierbei um einen Klassifikationsfall („Ja" / „Nein"-Response) handelt. Aus der angegebenen Formel ergibt sich mit den eingeführten Abkürzungen:

$$G = \hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2) = p(1-p) + q(1-q) = 1 - p^2 - q^2$$

b) Für Feature A ergibt sich zunächst folgendes Ergebnis:




Für Feature B ergibt sich entsprechend:



Mit der Formel aus Teil a) ergibt sich dann für das linke Blatt von Feature 1:

$$G_{1L} = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$$

Analog ergibt sich für das rechte Blatt:

$$G_{1R} = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

Gemittelt mit den Personen pro Blatt ergibt sich daraus:

$$TG_1 = \frac{5}{7}0{,}48 + \frac{2}{7}0.5 = 0.4857$$

Analog ergeben sich die Ergebnisse für Feature 2:

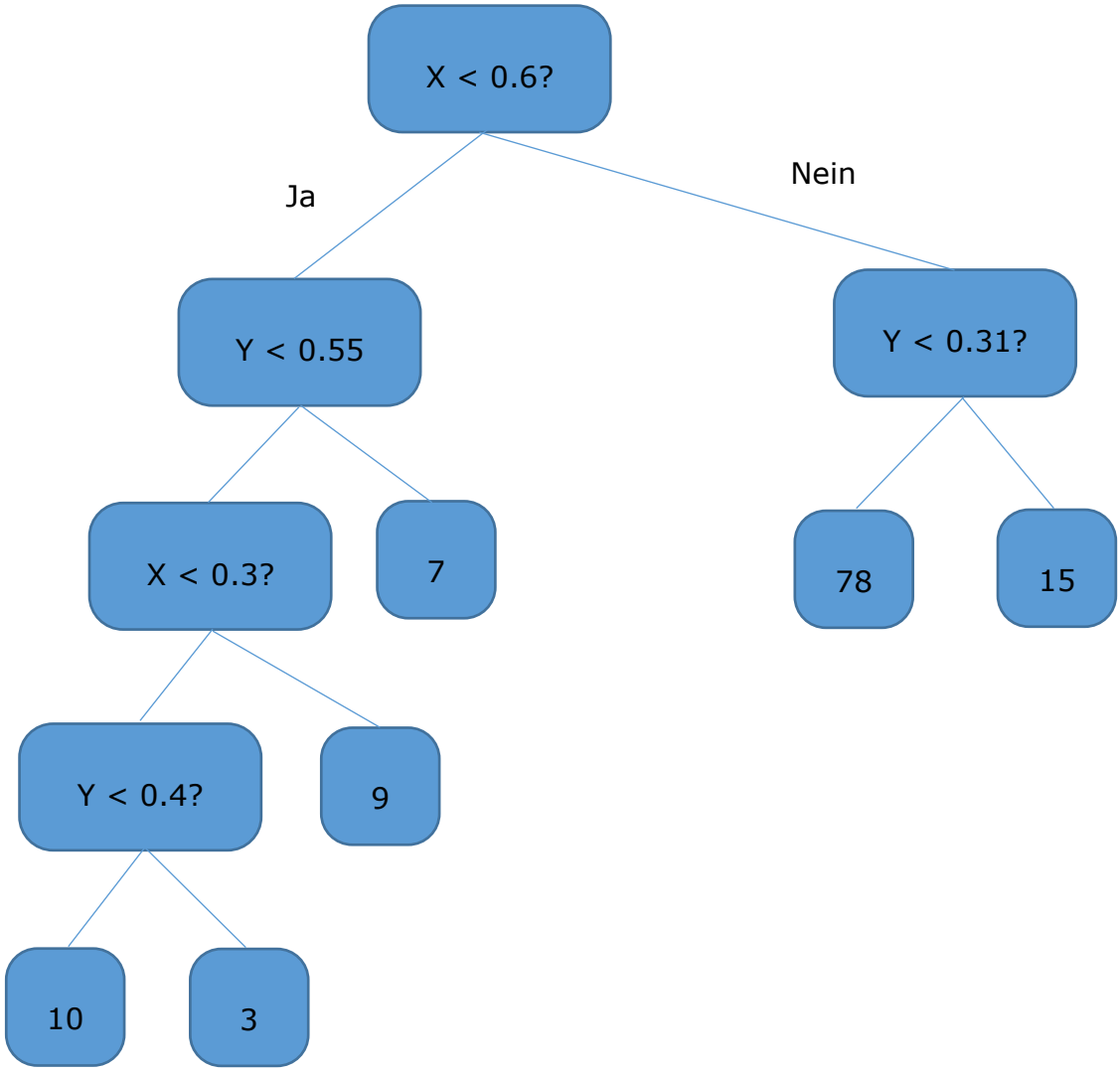
$$G_{2L} = 0.44, G_{2R} = 0.375$$

und daraus

$$TG_2 = 0.4029$$

Da $TG_2 < TG_1$, wird der Root Node mit Feature 2 versehen.

c) Aus der dargestellten Graphik ergibt sich der folgende Entscheidungsbaum:

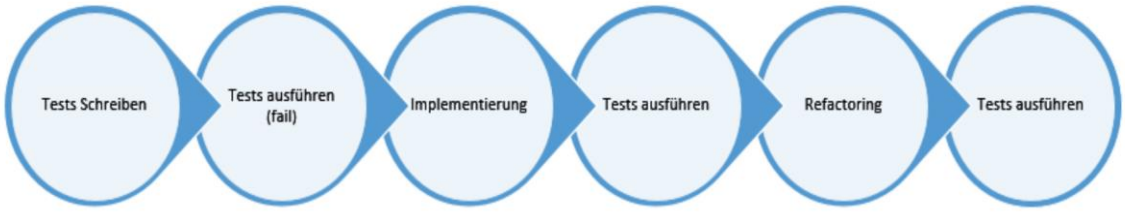


d) Skalierbarkeit bezeichnet die Fähigkeit eines Systems oder Algorithmus, auf die (erheblich) Veränderung einer Eingabegröße „angemessen" zu reagieren. Dabei kann unterschieden werden zwischen:

   a. Vertikaler Skalierbarkeit: Ausbau der verfügbaren Rechenleistung durch schnellere CPUs, mehr Haupt- und Massenspeicher, eine schnellere Netzwerkanbindung etc.

   b. Horizontale Skalierbarkeit: Skalierung durch Hinzufügen weiterer Rechner bzw. Knoten. Grundsätzlich kann eine beliebige Anzahl von weiteren Einheiten dem System hinzugefügt werden, allerdings sind i.A. Änderungen an den Algorithmen erforderlich, damit die zusätzlich zur Verfügung gestellten Ressourcen auch tatsächlich genutzt werden können (Parallelisierbarkeit).

   Der Prozess der Softwareentwicklung stellt sich wie folgt dar:




   Zunächst werden die Testfälle geschrieben (die vorerst fehlschlagen). Dann wird die Funktionalität programmiert, bis die Tests erfolgreich sind. Anschließend wird optimiert, wobei danach die Tests immer noch erfolgreich sein müssen.