Musterprüfung

Ausgabe 202310

Inhalt

Einführung

Dies ist die EXIN Data Analytics Foundation (DAF.DE) Musterprüfung. Es gilt die Prüfungsordnung von EXIN.

Die Musterprüfung besteht aus 40 Multiple-Choice-Fragen. Zu jeder Multiple-Choice-Frage werden mehrere Antwortmöglichkeiten angeboten. Es gibt jeweils eine richtige Antwort.

Sie können maximal 40 Punkte erreichen. Jede richtige Antwort zählt 1 Punkt. Um die Prüfung zu bestehen, müssen Sie mindestens 26 Punkte erzielen.

Die Bearbeitungszeit beträgt 60 Minuten.

Viel Erfolg!

Musterprüfung

1 / 40
Datenanalytik ist die Wissenschaft der Analyse von Daten, um daraus Schlussfolgerungen zu ziehen.

Was ist **kein** Fokus der Datenanalytik?

A) Die Umwandlung von Daten in Informationen
B) Die Datenerstellung
C) Die Datenpräsentation
D) Die Datenverarbeitung

2 / 40
Was ist ein Beispiel für einen Schritt der Datenbereinigung?

A) Ein Unternehmen sammelt die Umsatzerlöse aller Niederlassungen und stellt die Währungseinheit auf „tausend Dollar", bevor es die Datensätze in seinem Data Warehouse zusammenfasst.
B) Ein Unternehmen sammelt die Umsatzerlöse aller Niederlassungen und erstellt ein Datenmodell, um die Umsatzerlöse für das kommende Jahr nach Standorten vorherzusagen.
C) Ein Unternehmen sammelt die Umsatzerlöse aller Niederlassungen und löscht alle eventuellen Dublette, bevor es die Datensätze in seinem Data Warehouse zusammenfasst.
D) Ein Unternehmen sammelt die Umsatzerlöse aller Niederlassungen, ermittelt den Medianwert und identifiziert die Niederlassung, deren Umsatzerlös den größten Unterschied zum Medianwert aufweist.

3 / 40
Die Einhaltung von Rechtsvorschriften ist ein sehr wichtiger Bereich. Organisationen müssen sich dessen bewusst sein, um Risiken zu vermeiden.

In welchem Bereich, bezogen auf Daten, spielt die Einhaltung der Rechtsvorschriften **keine** wichtige Rolle?

A) Datenanalytik
B) Datenhoheit
C) Datenschutz
D) Geistiges Eigentum

4 / 40
Was ist **kein** Merkmal von interaktiven Dashboards?

A) Sie ermöglichen Benutzern, Daten detailliert aufzuschlüsseln.
B) Sie können Daten aus mehreren Quellen kombinieren.
C) Sie können für Benutzer, die nicht technisch versiert sind, schwierig zu benutzen sein.
D) Sie bieten dem Management eine Was-wäre-wenn-Analyse.

5 / 40
Eine Organisation möchte aus vielen verschiedenen Webseiten Text sammeln.

Wie lassen sich öffentliche Daten **am besten** beschaffen?

A) Mittels alternativer Daten
B) Mittels interner Erfassungssysteme
C) Mittels Interviews und Experimenten
D) Mittels Web Scraping

6 / 40
Was ist ein Beispiel für einen Kanal zur Sammlung von Daten?

A) Kontinuierliche Daten
B) Abstraktion der Absicht
C) Interne Erfassungssysteme
D) Schlüssel-Wert-Datenbank
E) Relationales Datenbankmanagementsystem (RDMS)
F) Strukturierte Daten

7 / 40
Was ist ein Beispiel für Web Scraping?

A) Herunterladen einer CSV-Datei von einer staatlichen Website
B) Herunterladen von Daten mittels Parsen der Inhalte einer Website
C) Herunterladen von Daten über eine File-Transfer-Protocol-Verbindung (FTP-Verbindung)
D) Herunterladen von Excel-Dateien, die im Internet gefunden werden

8 / 40
Was spielt für die Einhaltung des Datenrechts bei der Sammlung von Daten eine wesentliche Rolle?

A) Die Durchführung von Transaktionen mit offenen Daten
B) Die Sicherstellung von Transparenz bei der Sammlung von Daten
C) Das Mining minimaler Mengen von öffentlichen Daten
D) Die Beschaffung personenbezogener Daten

9 / 40
Welcher Datentyp wird durch Volumen, Geschwindigkeit und Vielfalt gekennzeichnet?

A) Alternative Daten
B) Big Data
C) Strukturierte Daten
D) Unstrukturierte Daten

10 / 40
Welche Datenlösung hat Entscheidungsfindung als ausdrückliches Ziel?

A) Enterprise Data Warehouse (EDW)
B) Schlüssel-Wert-Datenbank
C) Relationales Datenbankmanagementsystem (RDMS)
D) Unstrukturierte Datensysteme

11 / 40
Wie unterscheidet sich die Nutzung eines Enterprise Data Warehouses (EDW) von der eines relationalen Datenbankmanagementsystems (RDMS)?

A) Ein EDW ist eine Lösung für den Routinebetrieb und die täglichen Transaktionen, der Zugriff auf ein RDMS dagegen erfolgt für vorab festgelegte Auswertungen und Dashboards.
B) Ein EDW nutzt Online Analytical Processing (OLAP), ein RDMS dagegen Online Transaction Processing (OLTP).
C) Ein RDMS ist ein Werkzeug, das für komplexe Auswertungen und zur Analyse genutzt wird, während ein EDW für Hochgeschwindigkeitstransaktionen in Echtzeit eingesetzt wird.
D) Ein RDMS basiert auf einem Modell der Schlüssel-Wert-Datenbank, während ein EDW auf einem Modell der Objektspeicherung beruht.

12 / 40
Was ist **kein** Merkmal von verteilten Dateisystemen?

A) Die Erweiterung der Speicher- und Servermöglichkeiten zu vergleichsweise niedrigen Kosten
B) Die gleichzeitige Verarbeitung von Datenpaketen über mehrere Server
C) Das einfache Speichern von Daten als großes binäres Datenobjekt (BLOB) ohne Festlegung eines Schemas (Übersicht)
D) Das Speichern von Daten, die in ihrem Originalformat keine ausdrückliche Verwendung haben

13 / 40
Ein Fahrzeug-Leasingunternehmen hat für sein Kerngeschäft eine Cloud-Lösung, mit der es Angebote erstellen, Fahrzeuge beschaffen, die Instandhaltung organisieren und die Kreditwürdigkeit überprüfen kann.

Das Unternehmen möchte nun für die Datenanalytik ebenfalls eine Cloud-Lösung nutzen, die vom selben Cloud-Provider gehostet wird.

Was ist **kein** Vorteil der für die Datenanalytik vorgeschlagenen Lösung?

A) Cloud-Lösungen sind günstiger als die Hardware- und Softwareumgebung sowie die IT-Mitarbeiter selbst vorzuhalten.
B) Der Import und Export riesiger Datensätze auf verschiedene Server wird damit überflüssig.
C) Die Lösung lässt sich parallel zum Betrieb des Fahrzeug-Leasingunternehmens nach oben skalieren.
D) Die Lösung lässt sich nach unten skalieren, wenn für Datenanalytik kein Bedarf mehr besteht.

14 / 40
In welcher Beziehung stehen eine unabhängige und eine abhängige Variable?

A) Die Änderung der abhängigen Variablen durch einen Data Scientist wirkt sich auf die unabhängige Variable aus.
B) Die Werte einer abhängigen Variablen spiegeln die Auswirkungen wider, die nach der Änderung einer unabhängigen Variablen beobachtet und aufgezeichnet werden.
C) Eine unabhängige Variable wird durch ein kleines ‚y' und eine abhängige Variable als großes ‚X' wiedergegeben.
D) Eine unabhängige Variable sollte stets als kategorische Variable beschrieben werden, während eine abhängige Variable jede beliebige Form annehmen kann.

15 / 40
Um welchen Variablentyp handelt es sich bei einer kategorischen Variablen?

A) Boolesche Variable
B) Diskrete Variable
C) Nominale Variable
D) Numerische Variable

16 / 40
Um welchen Variablentyp handelt es sich beim Kilometerstand eines Fahrzeugs zum Jahresbeginn und zum Jahresende?

A) Boolesche
B) Kategorische
C) Numerische
D) Zeit-/Datum

17 / 40
Warum ist es wichtig, zwischen diskreten und kontinuierlichen Variablen zu unterscheiden?

A) Weil es hilft, die für die Variablen geeigneten Algorithmen auszuwählen
B) Weil es für die Festlegung des konzeptionellen Datenmodells wichtig ist
C) Weil nur diskrete Variablen mit anderen Variablen aggregiert werden können
D) Weil relationale Datenbanken nur kontinuierliche Variablen speichern können

18 / 40
Bei welcher Technik des Data Scrubbing werden Variablen in Integer umgewandelt?

A) One-Hot-Codierung
B) Zusammenführung von Variablen
C) Auswahl von Variablen
D) Web Scraping

19 / 40
Die Datenanalytik eines Fahrzeug-Leasingunternehmens basiert auf einem Datensatz der unter anderem folgende Spalten enthält: Vertragsnummer, Kennzeichen, Datum der Aufnahme, Marke, Typ, Farbe und Datum des Vertragsendes.

Die Spalten, Datum der Aufnahme' und ‚Datum des Vertragsendes' sind auf Integers mit sechs Ziffern reduziert. Die ersten vier Ziffern stehen dabei für das Jahr und die letzten zwei Ziffern für den Monat. Damit können die Verträge nach dem Monat des Vertragsendes in Kategorien aufgeteilt werden.

Für was ist dies ein Beispiel?

A) Klassenbildung
B) Zusammenführung von Variablen
C) One-Hot-Codierung
D) Auswahl von Variablen

20 / 40
Die Bewahrung der ursprünglichen, für das Data Scrubbing verwendeten Quelldaten ist wichtig, um Revisionen zu ermöglichen.

Was sollte genutzt werden, um die Bewahrung der Quelldaten sicherzustellen?

A) Eine Richtlinie zur Datenhaltung (Datenspeicherung)
B) Ein Data Warehouse
C) Eine Schlüssel-Wert-Datenbank
D) Algorithmen der künstlichen Intelligenz (KI)

21 / 40
Was ist ein Attribut von inferentiellen Methoden?

A) Sie eignen sich für Situationen, in denen Daten gut dokumentiert und in einem einheitlichen Informationspool standardisiert sind.
B) Sie fassen offensichtliche Trends zusammen und helfen so, komplexe Informationen in ein praktisches und leicht zu lesendes Format zu verdichten.
C) Sie isolieren und analysieren einen Teil der Daten und testen die Ergebnisse dann im Vergleich mit einer anderen Datenteilmenge.
D) Sie bieten eine praktische und bündige Zusammenfassung der historischen Daten, die aus einer potenziell riesigen Zahl von Einzelereignissen generiert wurden.

22 / 40
In welchem Fall sollte eine deskriptive Analyse genutzt werden?

A) Wenn ein Datensatz nicht leicht zu analysieren ist
B) Wenn nur begrenzte Informationen zur Verfügung stehen
C) Wenn die abgerufenen Daten sehr detailliert sind
D) Wenn die Aufzeichnungen Lücken enthalten

23 / 40
Was ist notwendig, damit ein menschlicher Operator mit der Data Mining (Datenauswertung) beginnt?

A) Eine als „wahrscheinlich“ oder „unwahrscheinlich“ definierte Hypothese
B) Eine große Zahl an möglichen Eingabekombinationen
C) Ein Modell, das enthält, welche Eingaben eine bestimmte Ausgabe erzeugen
D) Ein Problem, das als lösenswert erachtet wird

24 / 40
Welche Methode des maschinellen Lernens nutzt Trainings- und Testdaten?

A) Die Methode der Data Mining (Datenauswertung)
B) Die deskriptive Methode
C) Die inferentielle Methode
D) Die Methode der Split-Validierung

25 / 40
Ein Fahrzeug-Leasingunternehmen hat in seinem Vertriebsprozess einen großen Datensatz gesammelt. Das Unternehmen nutzt zur Feinabstimmung seiner Angebote maschinelles Lernen. Dabei werden nur Daten aus dem Vertriebsprozess ohne zusätzliche Regeln in die Software eingegeben. Die Software sucht dann nach Mustern und stellt ein Modell bereit, das das Angebot ohne zu große Ermäßigung berechnet.

Um welche Art von maschinellem Lernen handelt es sich?

A) Berstärkendes Lernen
B) Überwachtes Lernen
C) Unüberwachtes Lernen

26 / 40
Die Syntaxanalyse ist ein wichtiger Aspekt der natürlichen Sprachverarbeitung (NLP).

Was analysiert die Syntaxanalyse?

A) Die Bedeutung eines Satzes
B) Die benannten Entitäten
C) Den Text von Suchanfragen
D) Die Satzstruktur

27 / 40
Eine Aufgabe der natürlichen Sprachverarbeitung (NLP) umfasst die Identifizierung wichtiger, im Text erwähnter Entitäten, wie Name, Standort oder eine Aktivität.

Um welche NLP-Aufgabe handelt es sich?

A) Die Identifizierung von Klassen
B) Die Erkennung benannter Entitäten
C) Gefühlsanalyse
D) Stemming

28 / 40
Wie sind Daten und Algorithmen miteinander verbunden?

A) Algorithmen werden zur Verarbeitung und Analyse von Daten entwickelt.
B) Daten werden zur Umwandlung in algorithmisches Wissen entwickelt.
C) Verschiedene Personen nutzen bei der Verarbeitung der selben Daten die selben Algorithmen.
D) Die Eingabedaten ändern sich bei Algorithmen nie und führen so zur Eindeutigkeit.

29 / 40
Die Daten eines Immobilienmaklers zeigen, dass der durchschnittliche Verkaufspreis für Häuser in gleichem Maße sinkt, in dem die Hypothekenzinsen steigen.

Welche Art von Regressionsanalyse sollte genutzt werden, um dieses Muster in den Daten zu beschreiben?

A) Exponentielle Regressionsanalyse
B) Lineare Regressionsanalyse
C) Logistische Regressionsanalyse
D) Nichtlineare Regressionsanalyse

30 / 40
Welches Ziel verfolgt die Regressionsanalyse?

A) Die Unternehmensleistung, die in die Entscheidungsfindung einfließt, zu analysieren.
B) In den Fällen, in denen keine vorab festgelegten Klassen existieren, die Datenpunkte in Gruppen zu kategorisieren.
C) Eine Gerade oder eine Kurve zu finden, mit der sich die Muster in den Daten bestmöglich beschreiben lassen
D) Zu verstehen, wie Daten gesammelt, organisiert und interpretiert werden.

31 / 40
Welcher Modelltyp generiert Vorhersagen zu Kategorien und Clusterbildung?

A) Algorithmisches Modell
B) Klassifizierungsmodell
C) Regressionsmodell

32 / 40
Ein Unternehmen hat einen großen Datensatz über den Verkauf von Hemden. Die Mitarbeiter möchten die Zahlen besser verstehen und die Datenanalystin überlegt, ob eine Cluster-Analyse Erkenntnisse liefern könnte.

Inwiefern sorgt K-Means Clustering für ein besseres Verständnis der Verkaufszahlen?

A) Indem es die Daten in vorab festgelegte Gruppierungen einteilt
B) Indem es überraschende Beziehungen zwischen den Clustern beschreibt
C) Indem es die Rolle von Centroiden erklärt
D) Indem es einen neuen Ansatz zur Gruppierung von Kunden bereitstellt

33 / 40
Was ist ein **zentraler** Unterschied zwischen der Assoziationsanalyse und der Ablaufanalyse?

A) Die Assoziationsanalyse wendet die Technik des Generalized Sequential Patterns (GSP) an, während die Ablaufanalyse die Technik Apriori nutzt.
B) Die Assoziationsanalyse ignoriert die Reihenfolge, in der die Items erscheinen, die Ablaufanalyse dagegen berücksichtigt die Reihenfolge.
C) Die Assoziationsanalyse nutzt ein Konfidenzniveau zur Unterstützung der Entscheidungsfindung, aber bei der Ablaufanalyse dagegen handelt es sich um überwachtes Lernen.
D) Die Assoziationsanalyse funktioniert besser bei großen Datensätzen, während sich die Ablaufanalyse am besten für kleine Datensätze eignet.

34 / 40
Die Ablaufanalyse nutzt ein Konfidenzniveau für aufgeklärte Entscheidungen. Sie verwendet ferner eine Art von Kriterium, um sich auf häufige Muster zu konzentrieren und eine informierte Entscheidungsfindung zu gewährleisten.

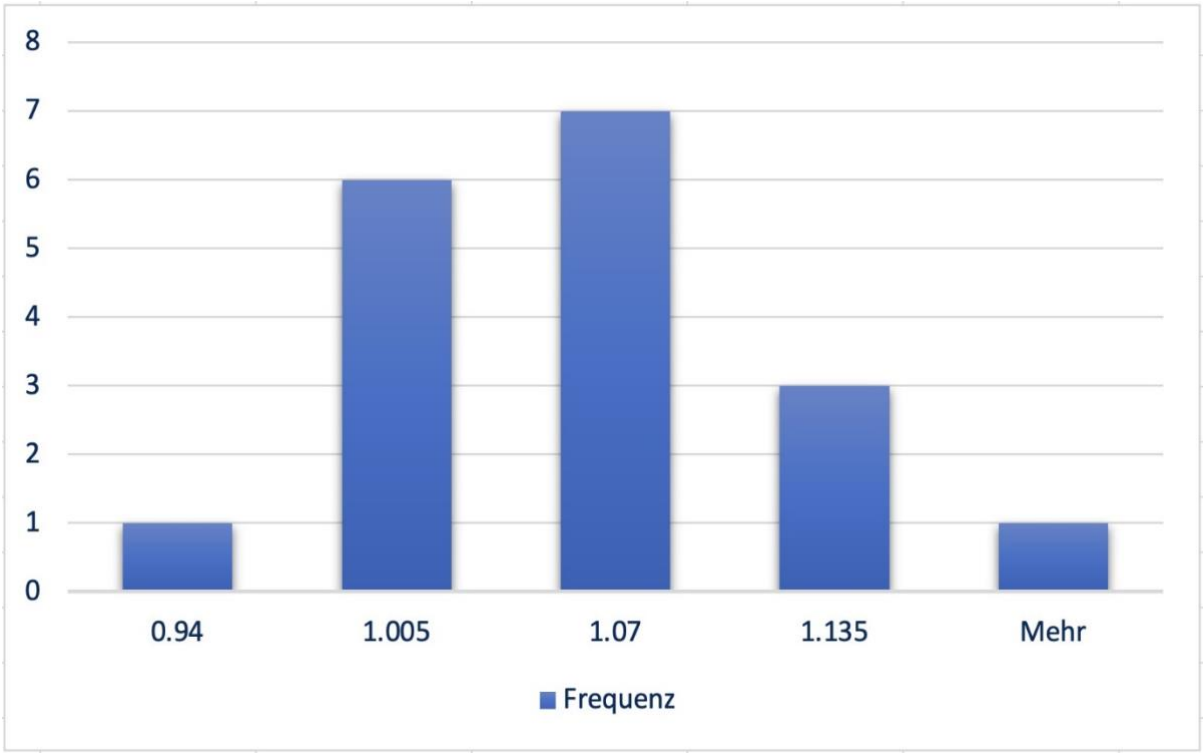
Um welches Kriterium handelt es sich?

A) Apriori-Algorithmus
B) Frequent Itemset
C) Mindest-Support
D) Rekursive Eliminierung

35 / 40
Welche grafische Technik wird in der Regel verwendet, wenn Daten einem externen Publikum bereitgestellt werden?

A) Expansiv
B) Erklärend
C) Erläuternd
D) Explorativ

36 / 40
Bitte betrachten Sie das Bild unten:



Für was ist dies ein Beispiel?

A) Balkendiagramm
B) Box-Plot
C) Heatmap
D) Histogramm

37 / 40
Welche Art von Plot (Ausdruck) wird verwendet, um die Streuung und Schiefe eines Datensatzes darzustellen?

A) Box-Plot
B) Rug-Plot
C) Scatter-Plot
D) Violin-Plot

38 / 40
Ein Immobilienmakler möchte eine Heatmap nutzen, um sich einen besseren Einblick über die Verkaufspreise von Häusern in den verschiedenen Stadtvierteln zu verschaffen.

Was zeigt die Heatmap?

A) Einen Stadtplan, auf dem die verschiedenen Preise der Häuser durch Farben dargestellt werden
B) Einen Stadtplan mit den verkauften Häusern und dem Verkaufspreis pro Haus
C) Eine Tabelle mit den durchschnittlichen Häuserpreisen pro Viertel, die im vergangenen Jahr erzielt wurden, sortiert nach dem Durchschnittspreis
D) Eine Tabelle mit den Preisen der im letzten Jahr verkauften Häuser geordnet nach Preisen

39 / 40
Warum ist ästhetische Gestaltung wichtig?

A) Sie verbessert die Techniken des maschinellen Lernens.
B) Sie erleichtert die natürliche Sprachverarbeitung (NLP).
C) Sie konzentriert sich auf Modelle der Regressionsanalyse.
D) Sie macht die Datenvisualisierung benutzerfreundlicher.

40 / 40
Welches Datenvisualisierungstool erinnert in Aussehen und Wirkung an Microsoft Excel?

A) Data Wrapper
B) Google Charts
C) Power BI
D) Tableau

Antwortschlüssel

1 / 40
Datenanalytik ist die Wissenschaft der Analyse von Daten, um daraus Schlussfolgerungen zu ziehen.

Was ist **kein** Fokus der Datenanalytik?

A) Die Umwandlung von Daten in Informationen
B) Die Datenerstellung
C) Die Datenpräsentation
D) Die Datenverarbeitung

A) Falsch. Das ist der Fokus der wichtigsten Aktivitäten im Bereich der Datenanalytik. Bei der Datenanalytik geht es darum, Daten zu analysieren und zu verarbeiten, um sich auf diese Weise Klarheit über diese Daten zu verschaffen.
B) Richtig. Zuerst werden die Daten erstellt und erst dann entsteht die Notwendigkeit, sich Klarheit über diese Daten zu verschaffen. Daher ist die Datenerstellung nicht Teil der Datenanalytik. (Literatur: B, Kapitel 2 und 3)
C) Falsch. Ein Fokus der Datenanalytik ist die Datenpräsentation, um sicherzustellen, dass die Ergebnisse der Analyse für das Zielpublikum geeignet sind.
D) Falsch. Wie Daten verarbeitet und analysiert werden, damit sie einen Sinn ergeben, ist eines der zentralen Themen der Datenanalytik.

2 / 40
Was ist ein Beispiel für einen Schritt der Datenbereinigung?

A) Ein Unternehmen sammelt die Umsatzerlöse aller Niederlassungen und stellt die Währungseinheit auf „tausend Dollar“, bevor es die Datensätze in seinem Data Warehouse zusammenfasst.
B) Ein Unternehmen sammelt die Umsatzerlöse aller Niederlassungen und erstellt ein Datenmodell, um die Umsatzerlöse für das kommende Jahr nach Standorten vorherzusagen.
C) Ein Unternehmen sammelt die Umsatzerlöse aller Niederlassungen und löscht alle eventuellen Dublette, bevor es die Datensätze in seinem Data Warehouse zusammenfasst.
D) Ein Unternehmen sammelt die Umsatzerlöse aller Niederlassungen, ermittelt den Medianwert und identifiziert die Niederlassung, deren Umsatzerlös den größten Unterschied zum Medianwert aufweist.

A) Falsch. Hierbei handelt es sich nicht um Datenbereinigung, sondern um Datennormalisierung/-standardisierung.
B) Falsch. Hierbei handelt es sich nicht um Datenbereinigung, sondern um prädiktive Datenanalyse.
C) Richtig. Das Löschen von Dubletten ist ein typischer Schritt der Datenbereinigung, der dazu dient, die Datenqualität beziehungsweise die Einmaligkeit der Daten sicherzustellen. (Literatur: B, Kapitel 4)
D) Falsch. Hierbei handelt es sich nicht um Datenbereinigung, sondern um Daten-Profiling. Dabei überprüft man Daten mit statistischen Funktionen, bevor man die Techniken des maschinellen Lernens anwendet.

3 / 40
Die Einhaltung von Rechtsvorschriften ist ein sehr wichtiger Bereich. Organisationen müssen sich dessen bewusst sein, um Risiken zu vermeiden.

In welchem Bereich, bezogen auf Daten, spielt die Einhaltung der Rechtsvorschriften **keine** wichtige Rolle?

A) Datenanalytik
B) Datenhoheit
C) Datenschutz
D) Geistiges Eigentum

A) Richtig. Die Gesetzgebung enthält keine Vorgaben dazu, wie Daten analysiert werden, sondern definiert lediglich, welche Daten für welchen Zweck gesammelt und analysiert werden dürfen. (Literatur: B, Kapitel 7)
B) Falsch. Immer mehr Länder verfügen aktuell über Vorschriften, die festlegen, wer für die gesammelten Daten verantwortlich ist und welche Rechte der Besitzer hat.
C) Falsch. Es gibt derzeit internationale Vorschriften zum Datenschutz, die für Organisationen durchaus sehr wichtig sind.
D) Falsch. Die Gesetzgebung zum geistigen Eigentum zählt zu den ersten Rechtsvorschriften, die sich auf die Datenanalytik auswirken.

4 / 40
Was ist **kein** Merkmal von interaktiven Dashboards?

A) Sie ermöglichen Benutzern, Daten detailliert aufzuschlüsseln.
B) Sie können Daten aus mehreren Quellen kombinieren.
C) Sie können für Benutzer, die nicht technisch versiert sind, schwierig zu benutzen sein.
D) Sie bieten dem Management eine Was-wäre-wenn-Analyse.

A) Falsch. Interaktive Dashboards ermöglichen den Benutzern durchaus eine detaillierte Aufschlüsselung der Daten.
B) Falsch. Interaktive Dashboards kombinieren Daten aus mehreren Quellen, um für ein umfassendes Verständnis zu sorgen.
C) Richtig. Interaktive Dashboards sind auf einfache Benutzung ausgelegt, auch von Bernutzern, die nicht technisch versiert sind. (Literatur: B, Kapitel 5)
D) Falsch. Interaktive Dashboards bieten durchaus Was-wäre-wenn-Erkenntnisse und erleichtern so die Entscheidungsfindung.

5 / 40
Eine Organisation möchte aus vielen verschiedenen Webseiten Text sammeln.

Wie lassen sich öffentliche Daten **am besten** beschaffen?

A) Mittels alternativer Daten
B) Mittels interner Erfassungssysteme
C) Mittels Interviews und Experimenten
D) Mittels Web Scraping

A) Falsch. Alternative Daten stellen Informationen aus nicht herkömmlichen Quellen zusammen, in der Regel Daten zu Banken und Finanzen. Daher ist dies nicht die beste Option.
B) Falsch. Die Beschaffung von Daten über interne Erfassungssysteme ist die gängigste Art, um interne Daten zu beschaffen.
C) Falsch. Interviews und Experimente eignen sich nicht für Automatisierung und Code. Außerdem sind Informationen aus Interviews keine öffentlichen Daten.
D) Richtig. Web Scraping sammelt Informationen mithilfe von Code und Automatisierung aus dem Internet. (Literatur: A, Kapitel 1)

6 / 40
Was ist ein Beispiel für einen Kanal zur Sammlung von Daten?

A) Kontinuierliche Daten
B) Abstraktion der Absicht
C) Interne Erfassungssysteme
D) Schlüssel-Wert-Datenbank
E) Relationales Datenbankmanagementsystem (RDMS)
F) Strukturierte Daten

A) Falsch. Kontinuierliche Daten sind ein Variablentyp, kein Kanal zur Sammlung von Daten.
B) Falsch. Die Absichtsabstraktion ist eine Komponente der natürlichen Sprachverarbeitung (NLP), kein Kanal zur Sammlung von Daten.
C) Richtig. Interne Erfassungssysteme sind ein Kanal zur Sammlung von Daten. (Literatur: A, Kapitel 1)
D) Falsch. Die Schlüssel-Wert-Datenbank ist ein System zur Speicherung von Daten, kein Kanal zur Sammlung von Daten.
E) Falsch. Das RDMS ist ein Software-Programm zur Speicherung von strukturierten Daten, kein Kanal zur Sammlung von Daten.
F) Falsch. Strukturierte Daten sind kein Kanal zur Sammlung von Daten, sondern ein Format für die Datenspeicherung.

7 / 40
Was ist ein Beispiel für Web Scraping?

A) Herunterladen einer CSV-Datei von einer staatlichen Website
B) Herunterladen von Daten mittels Parsen der Inhalte einer Website
C) Herunterladen von Daten über eine File-Transfer-Protocol-Verbindung (FTP-Verbindung)
D) Herunterladen von Excel-Dateien, die im Internet gefunden werden

A) Falsch. Web Scraping umfasst das Gewinnen wertvoller Daten. Dies wird durch die Kombination von Daten erreicht und nicht durch das Herunterladen einer einzelnen Datei.
B) Richtig. Beim Web Scraping werden Webseiten durchsucht und wertvolle Daten gewonnen. (Literatur: A, Kapitel 1)
C) Falsch. Hier fehlt, dass beim Web Scraping das Internet durchforstet wird.
D) Falsch. Zwar werden die Dateien im Internet gefunden, aber hier fehlt der Aspekt des Scraping.

8 / 40
Was spielt für die Einhaltung des Datenrechts bei der Sammlung von Daten eine wesentliche Rolle?

A) Die Durchführung von Transaktionen mit offenen Daten
B) Die Sicherstellung von Transparenz bei der Sammlung von Daten
C) Das Mining minimaler Mengen von öffentlichen Daten
D) Die Beschaffung personenbezogener Daten

A) Falsch. Offene Daten könnten persönliche Informationen enthalten. In diesem Fall würde das Datenrecht nicht unbedingt eingehalten.
B) Richtig. Die Transparenz, das heißt, darüber zu informieren, wie Daten gesammelt werden, spielt für die Einhaltung des geltenden Datenrechts eine wesentliche Rolle. (Literatur: A, Kapitel 1)
C) Falsch. Im Datenrecht wird nicht erwähnt, dass die Mining von öffentlichen Informationen eine wesentliche Rolle für die Einhaltung des Datenrechts spielt.
D) Falsch. Die Beschaffung personenbezogener Daten darf nur im Rahmen der gesetzlichen Vorgaben erfolgen und spielt daher für die Einhaltung des Datenrechts keine wesentliche Rolle.

9 / 40
Welcher Datentyp wird durch Volumen, Geschwindigkeit und Vielfalt gekennzeichnet?

A) Alternative Daten
B) Big Data
C) Strukturierte Daten
D) Unstrukturierte Daten

A) Falsch. Alternative Daten sind ein Kanal zur Datensammlung.
B) Richtig. Kennzeichnend für Big Data sind Volumen, Geschwindigkeit und Vielfalt. (Literatur: A, Kapitel 2)
C) Falsch. Strukturierte Daten stehen für Informationen, die in einem genau festgelegten und für Algorithmen einfachen Format organisiert sind, um wichtige Informationen leicht abrufen und verarbeiten zu können.
D) Falsch. Unstrukturierte Daten bestehen aus Informationen, die nicht erkennbar organisiert sind und nicht in ein standardisiertes Format wie einen tabellarischen Datensatz passen.

10 / 40
Welche Datenlösung hat Entscheidungsfindung als ausdrückliches Ziel?

A) Enterprise Data Warehouse (EDW)
B) Schlüssel-Wert-Datenbank
C) Relationales Datenbankmanagementsystem (RDMS)
D) Unstrukturierte Datensysteme

A) Richtig. Ausdrückliches Ziel eines EDW ist die Entscheidungsfindung nach der Analyse in Organisationen. (Literatur: A, Kapitel 2)
B) Falsch. In einer Schlüssel-Wert-Datenbank werden Daten als großes binäres Datenobjekt (BLOB) ohne Festlegung eines Schemas (Übersicht) gespeichert. Die Folgen sind ein vager Datenwert und dass sich das Ergebnis einer Anforderung nicht steuern lässt.
C) Falsch. Ein RDMS dient der Datenspeicherung, nicht der Entscheidungsfindung nach der Analyse.
D) Falsch. Unstrukturierte Datensysteme dienen in erster Linie der Speicherung und Verarbeitung von unstrukturierten Daten und werden nicht zur Entscheidungsfindung nach der Analyse genutzt.

11 / 40
Wie unterscheidet sich die Nutzung eines Enterprise Data Warehouses (EDW) von der eines relationalen Datenbankmanagementsystems (RDMS)?

A) Ein EDW ist eine Lösung für den Routinebetrieb und die täglichen Transaktionen, der Zugriff auf ein RDMS dagegen erfolgt für vorab festgelegte Auswertungen und Dashboards.
B) Ein EDW nutzt Online Analytical Processing (OLAP), ein RDMS dagegen Online Transaction Processing (OLTP).
C) Ein RDMS ist ein Werkzeug, das für komplexe Auswertungen und zur Analyse genutzt wird, während ein EDW für Hochgeschwindigkeitstransaktionen in Echtzeit eingesetzt wird.
D) Ein RDMS basiert auf einem Modell der Schlüssel-Wert-Datenbank, während ein EDW auf einem Modell der Objektspeicherung beruht.

A) Falsch. Ein EDW ist keine Lösung für den Routinebetrieb, denn es muss Daten aus verschiedenen Quellen konsolidieren und integrieren. Der Fokus eines EDW liegt auf analytischen Anwendungen. Ein RDMS ist ein Werkzeug zur Verwaltung von Transaktionsdaten.
B) Richtig. Ein EDW basiert auf OLAP, während ein RDMS auf OLTP basiert. (Literatur: A, Kapitel 2)
C) Falsch. Ein RDMS wird für Hochgeschwindigkeitstransaktionen in Echtzeit genutzt, ein EDW dagegen wird für komplexe Auswertungen und Analysen eingesetzt.
D) Falsch. Ein RDMS basiert nicht auf einem Modell einer Schlüssel-Wert-Datenbank, sondern auf einem relationalen Modell.

12 / 40
Was ist **kein** Merkmal von verteilten Dateisystemen?

A) Die Erweiterung der Speicher- und Servermöglichkeiten zu vergleichsweise niedrigen Kosten
B) Die gleichzeitige Verarbeitung von Datenpaketen über mehrere Server
C) Das einfache Speichern von Daten als großes binäres Datenobjekt (BLOB) ohne Festlegung eines Schemas (Übersicht)
D) Das Speichern von Daten, die in ihrem Originalformat keine ausdrückliche Verwendung haben

A) Falsch. Die Kosten für inkrementelle Speicherung und Server sind bei verteilten Dateisystemen im Vergleich niedriger.
B) Falsch. In verteilten Dateisystemen kann man Aufgaben im Netzwerk aufteilen und so Datenpakete gleichzeitig über verschiedene Server verarbeiten.
C) Richtig. In einer Schlüssel-Wert-Datenbank können Daten ohne Festlegung eines Schemas als BLOB gespeichert werden. Hierbei handelt es sich nicht um einen Vorteil von verteilten Dateisystemen. (Literatur: A, Kapitel 2)
D) Falsch. Verteilte Dateisysteme ermöglichen durchaus die Speicherung von Daten, die in ihrem Originalformat keine ausdrückliche Verwendung haben.

13 / 40
Ein Fahrzeug-Leasingunternehmen hat für sein Kerngeschäft eine Cloud-Lösung, mit der es Angebote erstellen, Fahrzeuge beschaffen, die Instandhaltung organisieren und die Kreditwürdigkeit überprüfen kann.

Das Unternehmen möchte nun für die Datenanalytik ebenfalls eine Cloud-Lösung nutzen, die vom selben Cloud-Provider gehostet wird.

Was ist **kein** Vorteil der für die Datenanalytik vorgeschlagenen Lösung?

A) Cloud-Lösungen sind günstiger als die Hardware- und Softwareumgebung sowie die IT-Mitarbeiter selbst vorzuhalten.
B) Der Import und Export riesiger Datensätze auf verschiedene Server wird damit überflüssig.
C) Die Lösung lässt sich parallel zum Betrieb des Fahrzeug-Leasingunternehmens nach oben skalieren.
D) Die Lösung lässt sich nach unten skalieren, wenn für Datenanalytik kein Bedarf mehr besteht.

A) Richtig. Die ist nicht immer der Fall. Eine Cloud-Lösung kann sogar kostspieliger sein. (Literatur: A, Kapitel 2)
B) Falsch. Alle Daten befinden sich bereits in der Cloud und können genutzt werden.
C) Falsch. Cloud-Lösungen können im Allgemeinen ohne Aufwand oder Investitionen seitens des Kunden nach oben skaliert werden.
D) Falsch. Ein Vorteil von Cloud-Lösungen ist, dass sie sich in der Regel problemlos nach oben und unten skalieren lassen.

14 / 40
In welcher Beziehung stehen eine unabhängige und eine abhängige Variable?

A) Die Änderung der abhängigen Variablen durch einen Data Scientist wirkt sich auf die unabhängige Variable aus.
B) Die Werte einer abhängigen Variablen spiegeln die Auswirkungen wider, die nach der Änderung einer unabhängigen Variablen beobachtet und aufgezeichnet werden.
C) Eine unabhängige Variable wird durch ein kleines ‚y' und eine abhängige Variable als großes ‚X' wiedergegeben.
D) Eine unabhängige Variable sollte stets als kategorische Variable beschrieben werden, während eine abhängige Variable jede beliebige Form annehmen kann.

A) Falsch. Die unabhängige Variable wird geändert und diese Änderung wirkt sich dann auf die abhängige Variable aus.
B) Richtig. Eine Änderung der unabhängigen Variablen, wirkt sich auf die abhängige Variable aus. (Literatur: A, Kapitel 2)
C) Falsch. Die unabhängige Variable wird in der Gleichung als ‚X' wiedergegeben.
D) Falsch. Eine unabhängige Variable kann nicht nur als kategorische Variable, sondern auch als numerische, boolesche oder Zeit-/Datumvariable beschrieben werden.

15 / 40
Um welchen Variablentyp handelt es sich bei einer kategorischen Variablen?

A) Boolesche Variable
B) Diskrete Variable
C) Nominale Variable
D) Numerische Variable

A) Falsch. Boolesche Variablen sind entweder wahr oder falsch und werden als 16-Bit-Werte (2-Byte-Werte) gespeichert. Eine Boolesche Variable ist ein anderer Variablentyp als eine kategorische Variable.
B) Falsch. Eine diskrete Variable nimmt eindeutige, zählbare Werte an. Kategorische Variablen dagegen sind nicht zählbar.
C) Richtig. Eine nominale Variable ist eine Art kategorische Variable. Es handelt sich dabei um eine Variable, die einen Wert aus einer begrenzten und in der Regel festen Zahl von möglichen Werten annehmen kann. Die werten sind qualitativ in Natur und können nicht aggregiert worden. (Literatur: A, Kapitel 3)
D) Falsch. Numerische Variablen haben Werte, die eine messbare Quantität als Zahl beschreiben, wie zum Beispiel ‚wie viele' oder ‚wie viel'. Kategorische Variablen sind nicht zählbar.

16 / 40
Um welchen Variablentyp handelt es sich beim Kilometerstand eines Fahrzeugs zum Jahresbeginn und zum Jahresende?

A) Boolesche
B) Kategorische
C) Numerische
D) Zeit-/Datum

A) Falsch. Der Kilometerstand kann zwar über oder unter dem Jahreslimit liegen, aber der Wert selbst ist kein boolescher Wert, sondern eine beliebige Zahl größer als 0.
B) Falsch. Die Kilometerstandvariable kann zwar mit einer kategorischen Variablen (den in einem bestimmten Leasingvertrag festgelegten Limits) verglichen werden, beim Kilometerstand selbst jedoch handelt es sich nicht um eine kategorische Variable.
C) Richtig. Die Variable lässt sich mathematisch als Integer ausdrücken. (Literatur: A, Kapitel 3)
D) Falsch. Die gefahrenen Kilometer sind zwar mit einem bestimmten Zeitraum verbunden, dies gilt aber nicht für die Zahl an sich.

17 / 40
Warum ist es wichtig, zwischen diskreten und kontinuierlichen Variablen zu unterscheiden?

A) Weil es hilft, die für die Variablen geeigneten Algorithmen auszuwählen
B) Weil es für die Festlegung des konzeptionellen Datenmodells wichtig ist
C) Weil nur diskrete Variablen mit anderen Variablen aggregiert werden können
D) Weil relationale Datenbanken nur kontinuierliche Variablen speichern können

A) Richtig. Zu erkennen, ob es sich bei einer Variablen um eine diskrete oder eine kontinuierliche Variable handelt, ist für die Analyse von Daten wichtig, denn es bestimmt, ob eine Variable mit dem gewählten Algorithmus kompatibel ist oder nicht. (Literatur: A, Kapitel 2)
B) Falsch. Das Erkennen kontinuierlicher oder diskreter Variablen beeinträchtigt das konzeptionelle Datenmodell nicht. Dieses hält die höheren Abstraktionsinhalte (Entitäten und Beziehungen) fest.
C) Falsch. Diskrete Variablen können nicht mit anderen Variablen aggregiert oder mathematisch manipuliert werden.
D) Falsch. Relationale Datenbanken können sowohl diskrete als auch kontinuierliche Variablen speichern.

18 / 40
Bei welcher Technik des Data Scrubbing werden Variablen in Integer umgewandelt?

A) One-Hot-Codierung
B) Zusammenführung von Variablen
C) Auswahl von Variablen
D) Web Scraping

A) Richtig. Bei der Technik der One-Hot-Codierung werden Variablen zur erfolgreichen Ausführung von Algorithmen als Integer dargestellt. (Literatur: A, Kapitel 4)
B) Falsch. Bei der Zusammenführung von Variablen werden verbundene Variablen kombiniert, um ein Maximum an Informationen zu bewahren.
C) Falsch. Bei der Auswahl von Variablen werden aus einem Datensatz die Variablen gewählt, die wertvoller sind als andere.
D) Falsch. Web Scraping ist keine Technik des Data Scrubbing, sondern eine Methode zur Datensammlung.

19 / 40
Die Datenanalytik eines Fahrzeug-Leasingunternehmens basiert auf einem Datensatz der unter anderem folgende Spalten enthält: Vertragsnummer, Kennzeichen, Datum der Aufnahme, Marke, Typ, Farbe und Datum des Vertragsendes.

Die Spalten, Datum der Aufnahme' und ‚Datum des Vertragsendes' sind auf Integers mit sechs Ziffern reduziert. Die ersten vier Ziffern stehen dabei für das Jahr und die letzten zwei Ziffern für den Monat. Damit können die Verträge nach dem Monat des Vertragsendes in Kategorien aufgeteilt werden.

Für was ist dies ein Beispiel?

A) Klassenbildung
B) Zusammenführung von Variablen
C) One-Hot-Codierung
D) Auswahl von Variablen

A) Richtig. Bei der Klassenbildung werden numerische Werte oder Zeitstempelwerte mit Hilfe eines diskreten Integers in eine Kategorie umgewandelt. (Literatur: A, Kapitel 4)
B) Falsch. Eine Zusammenführung der zwei Variablen ‚Datum der Hereinnahme' und ‚Datum des Vertragsendes' findet nicht statt.
C) Falsch. Bei der One-Hot-Codierung werden diskrete Variablen in ein binäres Format überführt. Die sich daraus ergebenden Integer können mehr als zwei Werte umfassen und sind keine Zustände.
D) Falsch. Bei der Auswahl von Variablen geht es darum, Variablen auszulassen, die keine Erkenntnisse bieten. Dies ist in diesem Beispiel aber nicht der Fall.

20 / 40
Die Bewahrung der ursprünglichen, für das Data Scrubbing verwendeten Quelldaten ist wichtig, um Revisionen zu ermöglichen.

Was sollte genutzt werden, um die Bewahrung der Quelldaten sicherzustellen?

A) Eine Richtlinie zur Datenhaltung (Datenspeicherung)
B) Ein Data Warehouse
C) Eine Schlüssel-Wert-Datenbank
D) Algorithmen der künstlichen Intelligenz (KI)

A) Richtig. Die Erstellung und Nutzung einer Richtlinie zur Datenhaltung ist wichtig, um sicherzustellen, dass die Quelldaten bewahrt werden. (Literatur: A, Kapitel 4)
B) Falsch. Ein Data Warehouse ist eine relationale Datenbank. Diese speichert Daten, stellt aber nicht sicher, dass die Quelldaten bewahrt bleiben. Dies sollte mittels Data Governance und den Regeln der Datenverwaltung gewährleistet werden.
C) Falsch. Eine Schlüssel-Wert-Datenbank ist eine Datenbank, in der Daten als Schlüssel-Wert-Paar gespeichert werden.
D) Falsch. KI-Algorithmen sorgen nicht dafür, dass die ursprünglichen Quelldaten bewahrt bleiben. KI wird für Erkenntnisse, Vorhersagen sowie Inferenz eingesetzt.

21 / 40
Was ist ein Attribut von inferentiellen Methoden?

A) Sie eignen sich für Situationen, in denen Daten gut dokumentiert und in einem einheitlichen Informationspool standardisiert sind.
B) Sie fassen offensichtliche Trends zusammen und helfen so, komplexe Informationen in ein praktisches und leicht zu lesendes Format zu verdichten.
C) Sie isolieren und analysieren einen Teil der Daten und testen die Ergebnisse dann im Vergleich mit einer anderen Datenteilmenge.
D) Sie bieten eine praktische und bündige Zusammenfassung der historischen Daten, die aus einer potenziell riesigen Zahl von Einzelereignissen generiert wurden.

A) Falsch. Dies ist ein Merkmal der deskriptiven Analyse und kein Attribut von inferentiellen Methoden.
B) Falsch. Dies ist ein Merkmal der deskriptiven Analyse. Inferentielle Methoden werden nicht zur Verdichtung von Informationen genutzt.
C) Richtig. Dies beschreibt die Split-Validierung, eine Technik, die bei inferentiellen Methoden zum Einsatz kommt. (Literatur: A, Kapitel 5)
D) Falsch. Dies bezieht sich auf die deskriptive Analyse. Inferentielle Methoden befassen sich mit der natürlichen Varianz relativ nicht so bündiger Daten.

22 / 40
In welchem Fall sollte eine deskriptive Analyse genutzt werden?

A) Wenn ein Datensatz nicht leicht zu analysieren ist
B) Wenn nur begrenzte Informationen zur Verfügung stehen
C) Wenn die abgerufenen Daten sehr detailliert sind
D) Wenn die Aufzeichnungen Lücken enthalten

A) Falsch. In diesem Fall sollte die inferentielle statistische Analyse genutzt werden.
B) Falsch. In diesem Fall sollte die inferentielle statistische Analyse genutzt werden.
C) Richtig. Die deskriptive Analyse sollte genutzt werden, wenn detaillierte Daten zur Verfügung stehen. (Literatur: A, Kapitel 5)
D) Falsch. In diesem Fall sollte die inferentielle statistische Analyse genutzt werden.

23 / 40
Was ist notwendig, damit ein menschlicher Operator mit der Data Mining (Datenauswertung) beginnt?

A) Eine als „wahrscheinlich“ oder „unwahrscheinlich“ definierte Hypothese
B) Eine große Zahl an möglichen Eingabekombinationen
C) Ein Modell, das enthält, welche Eingaben eine bestimmte Ausgabe erzeugen
D) Ein Problem, das als lösenswert erachtet wird

A) Falsch. Dies benötigt ein Operator für ein statistisches Standardmodell.
B) Falsch. Dies benötigt ein Operator für das berstärkende Lernen.
C) Falsch. Dies benötigt ein Operator für das überwachte Lernen.
D) Richtig. Ein Operator braucht ein Ziel, wie zum Beispiel ein Problem, das als lösenswert erachtet wird. (Literatur: A, Kapitel 5)

24 / 40
Welche Methode des maschinellen Lernens nutzt Trainings- und Testdaten?

A) Die Methode der Data Mining (Datenauswertung)
B) Die deskriptive Methode
C) Die inferentielle Methode
D) Die Methode der Split-Validierung

A) Falsch. Das Data Mining konzentriert sich darauf, Neues zu lernen. Das maschinelle Lernen dagegen konzentriert sich darauf, wie man das Gelernte in einem bestimmten Kontext anwenden kann.
B) Falsch. Deskriptive Analyse eignet sich besonders für Fälle, in denen Daten gut dokumentiert und in einem einheitlichen Informationspool standardisiert werden.
C) Richtig. Inferentielle Methoden nutzen 70 bis 80 % eines Datensatzes, um ein Analysewerkzeug darin zu schulen, Muster zu erkennen und diese an den verbleibenden 20 bis 30% des Datensatzes zu testen. (Literatur: A, Kapitel 5)
D) Falsch. Inferentielle Methoden nutzen häufig eine Technik, die als Split-Validierung bezeichnet wird. Dabei handelt es sich jedoch nicht um die gesamte Methode.

25 / 40
Ein Fahrzeug-Leasingunternehmen hat in seinem Vertriebsprozess einen großen Datensatz gesammelt. Das Unternehmen nutzt zur Feinabstimmung seiner Angebote maschinelles Lernen. Dabei werden nur Daten aus dem Vertriebsprozess ohne zusätzliche Regeln in die Software eingegeben. Die Software sucht dann nach Mustern und stellt ein Modell bereit, das das Angebot ohne zu große Ermäßigung berechnet.

Um welche Art von maschinellem Lernen handelt es sich?

A) Berstärkendes Lernen
B) Überwachtes Lernen
C) Unüberwachtes Lernen

A) Falsch. Zwar gibt es eine große Zahl möglicher Eingabekombinationen, aber deren Leistung wird nicht nach dem Zufallsprinzip benotet.
B) Falsch. Es gibt keine spezifische oder präzise Regel, um den korrekten Preis auf der Basis von unabhängigen Eingabevariablen zu berechnen.
C) Richtig. Zwar gibt es vorab gekennzeichnete Eingabe- und Ausgabekombinationen, aber es gibt keine bekannten Ausgaben, die als Referenzpunkte dienen können. Unüberwachtes Lernen analysiert die Eingabewerte und sucht so nach Mustern für die bestmögliche Preisgestaltung. (Literatur: A, Kapitel 5)

26 / 40
Die Syntaxanalyse ist ein wichtiger Aspekt der natürlichen Sprachverarbeitung (NLP).

Was analysiert die Syntaxanalyse?

A) Die Bedeutung eines Satzes
B) Die benannten Entitäten
C) Den Text von Suchanfragen
D) Die Satzstruktur

A) Falsch. Dies nennt man Semantikanalyse.
B) Falsch. Dies nennt man Klassifizierung.
C) Falsch. Dies nennt man Textparsen.
D) Richtig. Die Analyse der Satzstruktur nennt man Syntaxanalyse. (Literatur: A, Kapitel 10)

27 / 40
Eine Aufgabe der natürlichen Sprachverarbeitung (NLP) umfasst die Identifizierung wichtiger, im Text erwähnter Entitäten, wie Name, Standort oder eine Aktivität.

Um welche NLP-Aufgabe handelt es sich?

A) Die Identifizierung von Klassen
B) Die Erkennung benannter Entitäten
C) Gefühlsanalyse
D) Stemming

A) Falsch. Bei der Identifizierung von Klassen wird ein Dokument mit den binären Vektoren bestimmter Token gekennzeichnet.
B) Richtig. Bei der Erkennung benannter Entitäten werden die wichtigsten Teile eines Texts ausgewählt, beispielsweise das ‚Was', ‚Wo' und ‚Wie'. (Literatur: A, Kapitel 10)
C) Falsch. Die Gefühlsanalyse wird benutzt, um die emotionale Absicht eines Dokuments zu entdecken.
D) Falsch. Stemming wird benutzt, um die Schlüsselwörter eines Dokuments auf ihre Stammform zu parsen.

28 / 40
Wie sind Daten und Algorithmen miteinander verbunden?

A) Algorithmen werden zur Verarbeitung und Analyse von Daten entwickelt.
B) Daten werden zur Umwandlung in algorithmisches Wissen entwickelt.
C) Verschiedene Personen nutzen bei der Verarbeitung der selben Daten die selben Algorithmen.
D) Die Eingabedaten ändern sich bei Algorithmen nie und führen so zur Eindeutigkeit.

A) Richtig. Ein Algorithmus ist eine Schrittfolge, die auf die Hinweise und die sich ändernden Muster reagiert, die aus den Daten herrühren. (Literatur: A, Kapitel 6)
B) Falsch. Es besteht kein Zusammenhang zwischen der Datenentwicklung und algorithmischem Wissen.
C) Falsch. Verschiedene Personen nutzen zur Bewertung von Daten ihre eigenen internen Algorithmen.
D) Falsch. Daten ändern sich ständig und sind für Algorithmen nicht eindeutig.

29 / 40
Die Daten eines Immobilienmaklers zeigen, dass der durchschnittliche Verkaufspreis für Häuser in gleichem Maße sinkt, in dem die Hypothekenzinsen steigen.

Welche Art von Regressionsanalyse sollte genutzt werden, um dieses Muster in den Daten zu beschreiben?

A) Exponentielle Regressionsanalyse
B) Lineare Regressionsanalyse
C) Logistische Regressionsanalyse
D) Nichtlineare Regressionsanalyse

A) Falsch. Die exponentielle Regressionsanalyse eignet sich nicht bei linearen Beziehungen zwischen den Variablen.
B) Richtig. Die Regression ist linear und negativ. (Literatur: A, Kapitel 7)
C) Falsch. Die logistische Regression liefert diskrete Variablen als Ausgaben.
D) Falsch. Die Regression ist linear. Daher eignet sich eine lineare Regressionsanalyse in diesem Fall am besten.

30 / 40
Welches Ziel verfolgt die Regressionsanalyse?

A) Die Unternehmensleistung, die in die Entscheidungsfindung einfließt, zu analysieren.
B) In den Fällen, in denen keine vorab festgelegten Klassen existieren, die Datenpunkte in Gruppen zu kategorisieren.
C) Eine Gerade oder eine Kurve zu finden, mit der sich die Muster in den Daten bestmöglich beschreiben lassen
D) Zu verstehen, wie Daten gesammelt, organisiert und interpretiert werden.

A) Falsch. Dies bezieht sich auf die Geschäftsanalytik (BI). Diese kann als Werkzeugkasten definiert werden, um Informationen über die Leistung der Organisation zu sammeln, zu analysieren und an die Entscheidungsträger zu berichten.
B) Falsch. Die logistische Regression kann zwar genutzt werden, um Aufzeichnungen mit ähnlichen oder naheliegenden Werten zu identifizieren, dies trifft jedoch nicht auf alle Regressionsmethoden zu.
C) Richtig. Genau das ist das Ziel der Regressionsanalyse. Eine einzige Gerade oder Kurve kann zwar eine zu starke Vereinfachung der Daten darstellen, bietet aber einen nützlichen Referenzpunkt für allgemeine Vorhersagen über künftige Daten. (Literatur: A, Kapitel 7)
D) Falsch. Dies bezieht sich auf die Statistik. Deren primäres Ziel ist es, die Bedeutung von Daten und ihren Abweichungen zu bestimmen.

31 / 40
Welcher Modelltyp generiert Vorhersagen zu Kategorien und Clusterbildung?

A) Algorithmisches Modell
B) Klassifizierungsmodell
C) Regressionsmodell

A) Falsch. Algorithmen werden zwar in allen Techniken der Datenmodellierung genutzt, sind aber selbst kein Modell.
B) Richtig. Klassifizierung ist der Oberbegriff für Algorithmen, die für Themen wie Marktforschung und Kunden-Profiling Vorhersagen zu Kategorien generieren und die Cluster-Analyse analysieren. (Literatur: A, Kapitel 8)
C) Falsch. Die Regressionsanalyse ist eine populäre statistische Technik zur Modellierung der Beziehung zwischen einer oder mehreren unabhängigen und einer abhängigen Variablen.

32 / 40
Ein Unternehmen hat einen großen Datensatz über den Verkauf von Hemden. Die Mitarbeiter möchten die Zahlen besser verstehen und die Datenanalystin überlegt, ob eine Cluster-Analyse Erkenntnisse liefern könnte.

Inwiefern sorgt K-Means Clustering für ein besseres Verständnis der Verkaufszahlen?

A) Indem es die Daten in vorab festgelegte Gruppierungen einteilt
B) Indem es überraschende Beziehungen zwischen den Clustern beschreibt
C) Indem es die Rolle von Centroiden erklärt
D) Indem es einen neuen Ansatz zur Gruppierung von Kunden bereitstellt

A) Falsch. Die Nutzung von K-Means Clustering ermöglicht die Entdeckung bislang nicht identifizierter Gruppen.
B) Falsch. Die Nutzung von K-Means Clustering ermöglicht es, maximal voneinander getrennte Gruppe zu finden. Eine Analyse der Beziehung zwischen diesen Gruppen ist jedoch nicht möglich.
C) Falsch. Centroide sind eine technische Lösung für die Gruppenbildung, spielen jedoch keine bedeutende Rolle.
D) Richtig. K-Means Clustering erweist sich dann als nützlich, wenn keine bestehende Kategorie bekannt ist, aber neue, bislang unidentifizierte Gruppierungen gefunden werden sollen. Die Verwendung dieser Gruppierungen kann neue Erkenntnisse bieten. (Literatur: A, Kapitel 8)

33 / 40
Was ist ein **zentraler** Unterschied zwischen der Assoziationsanalyse und der Ablaufanalyse?

A) Die Assoziationsanalyse wendet die Technik des Generalized Sequential Patterns (GSP) an, während die Ablaufanalyse die Technik Apriori nutzt.
B) Die Assoziationsanalyse ignoriert die Reihenfolge, in der die Items erscheinen, die Ablaufanalyse dagegen berücksichtigt die Reihenfolge.
C) Die Assoziationsanalyse nutzt ein Konfidenzniveau zur Unterstützung der Entscheidungsfindung, aber bei der Ablaufanalyse dagegen handelt es sich um überwachtes Lernen.
D) Die Assoziationsanalyse funktioniert besser bei großen Datensätzen, während sich die Ablaufanalyse am besten für kleine Datensätze eignet.

A) Falsch. Die Ablaufanalyse nutzt GSP, die Assoziationsanalyse nutzt Apriori.
B) Richtig. Die Assoziationsanalyse ignoriert die Reihenfolge der Ereignisse; für die Ablaufanalyse ist diese jedoch wichtig. (Literatur: A, Kapitel 9)
C) Falsch. Bei der Ablaufanalyse handelt es sich nicht um überwachtes Lernen, sondern um eine Art von Assoziationsanalyse.
D) Falsch. Die Assoziationsanalyse und die Ablaufanalyse unterscheiden sich nicht bezüglich der Größe der Datensätze.

34 / 40
Die Ablaufanalyse nutzt ein Konfidenzniveau für aufgeklärte Entscheidungen. Sie verwendet ferner eine Art von Kriterium, um sich auf häufige Muster zu konzentrieren und eine informierte Entscheidungsfindung zu gewährleisten.

Um welches Kriterium handelt es sich?

A) Apriori-Algorithmus
B) Frequent Itemset
C) Mindest-Support
D) Rekursive Eliminierung

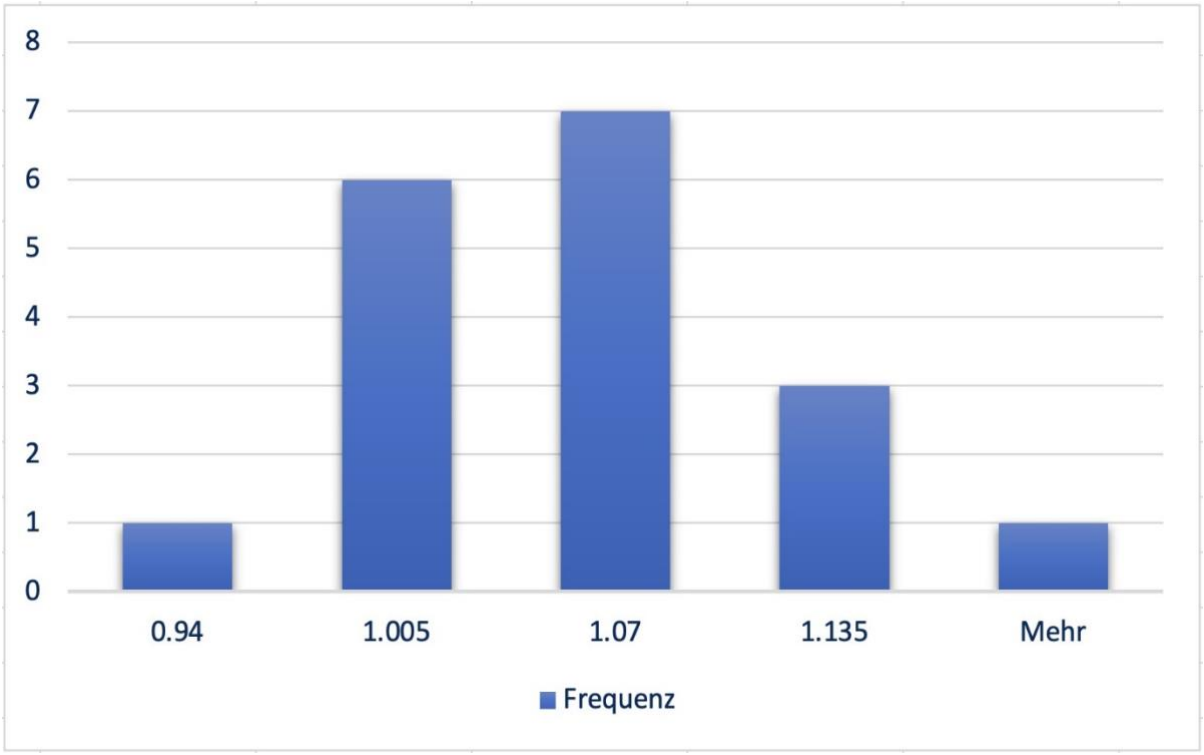
A) Falsch. Apriori ist ein Algorithmus für die Auswertung des Frequent Itemset und das Lernen von Assoziationsregeln über relationale Datenbanken.
B) Falsch. Ein Frequent Itemset tritt in Mindest-Support-Transaktionen aus der Transaktionsdatenbank auf, für die der Benutzer den Mindest-Support als Schwelle gesetzt hat.
C) Richtig. Die Ablaufanalyse nutzt ebenso wie die Assoziationsanalyse Mindest-Support-Kriterien, um sich auf häufige Muster zu konzentrieren, sowie ein Konfidenzniveau, um informierte Entscheidungen zu Gewährleistung und die Implementierung entsprechend zu priorisieren. (Literatur: A, Kapitel 10)
D) Falsch. Die rekursive Eliminierung ist ein Algorithmus zur Entdeckung von Frequent Item Sets in einer Transaktionsdatenbank.

35 / 40
Welche grafische Technik wird in der Regel verwendet, wenn Daten einem externen Publikum bereitgestellt werden?

A) Expansiv
B) Erklärend
C) Erläuternd
D) Explorativ

A) Falsch. Der Begriff „Expansiv“ wird in der Regel mit dem Wort „Politik“ kombiniert. Mit dem Begriff bezeichnet man in der Wirtschaft Techniken zur Verhinderung von Rezession und Arbeitslosigkeit.
B) Richtig. Erklärende Grafiken haben das Ziel, Daten und Erkenntnisse vereinfacht darzustellen, um dem Publikum zu helfen, komplexe Daten besser zu verstehen. (Literatur: A, Kapitel 11)
C) Falsch. Erläuternd ist eine philosophische und literarische Technik, die in der Datenanalytik nicht genutzt wird.
D) Falsch. Explorative Grafiken werden spontan erstellt, um, während die Analyse noch in Gange ist und sich das Modell noch im Produktionsmodus befindet, das unternehmensinterne Verstehen zu unterstützen.

36 / 40
Bitte betrachten Sie das Bild unten:



Für was ist dies ein Beispiel?

A) Balkendiagramm
B) Box-Plot
C) Heatmap
D) Histogramm

A) Richtig. Ein Balkendiagramm ist eine Grafik, bei der kategorische Daten mittels rechteckiger Balken dargestellt werden. Die Länge der einzelnen Balken ist dabei proportional zu dem Wert, den sie repräsentieren. (Literatur: A, Kapitel 12)
B) Falsch. Ein Box-Plot ist ein Diagramm, das die fünfstellige Zusammenfassung eines Datensatzes zeigt.
C) Falsch. Eine Heatmap ist eine zweidimensionale Darstellung von Daten, bei der die Werte durch Farben wiedergegeben werden.
D) Falsch. Ein Histogramm ist die grafische Darstellung von Daten, die in kontinuierliche Zahlenbereiche gruppiert sind, die vertikalen Balken entsprechen. In einem Histogramm gibt es keine Leerstellen zwischen den Balken.

37 / 40
Welche Art von Plot (Ausdruck) wird verwendet, um die Streuung und Schiefe eines Datensatzes darzustellen?

A) Box-Plot
B) Rug-Plot
C) Scatter-Plot
D) Violin-Plot

A) Richtig. Der Box-Plot beschreibt die Symmetrie der Daten und wird verwendet, um die Verteilung eines kontinuierlichen Datensatzes zusammenzufassen und darzustellen. (Literatur: A, Kapitel 11)
B) Falsch. Der Rug-Plot visualisiert die Verteilung einer einzigen Variablen, nicht die Schiefe eines Datensatzes.
C) Falsch. Der Scatter-Plot vermittelt Informationen über die Beziehung zwischen kontinuierlichen Variablen, nicht über die Verteilung und Schiefe eines Datensatzes.
D) Falsch. Der Violin-Plot beschreibt die Verteilung der Daten und ihre Dichte, jedoch nicht ihre Schiefe.

38 / 40
Ein Immobilienmakler möchte eine Heatmap nutzen, um sich einen besseren Einblick über die Verkaufspreise von Häusern in den verschiedenen Stadtvierteln zu verschaffen.

Was zeigt die Heatmap?

A) Einen Stadtplan, auf dem die verschiedenen Preise der Häuser durch Farben dargestellt werden
B) Einen Stadtplan mit den verkauften Häusern und dem Verkaufspreis pro Haus
C) Eine Tabelle mit den durchschnittlichen Häuserpreisen pro Viertel, die im vergangenen Jahr erzielt wurden, sortiert nach dem Durchschnittspreis
D) Eine Tabelle mit den Preisen der im letzten Jahr verkauften Häuser geordnet nach Preisen

A) Richtig. Eine Heatmap stellt Daten als Farben dar. (Literatur: A, Kapitel 12)
B) Falsch. Eine Heatmap stellt Daten als Farben dar. Bei diesem Plan gibt es keinen Hinweis, dass Farben genutzt werden.
C) Falsch. Eine Heatmap stellt Daten als Farben dar. Diese Tabelle ist keine Karte und es gibt keinen Hinweis, dass Farben genutzt werden.
D) Falsch. Eine Heatmap stellt Daten als Farben dar. Bei dieser Tabelle handelt es sich nicht um eine Karte und es gibt keinen Hinweis, dass Farben genutzt werden. Außerdem müssen die Preise bei einer Heatmap nicht in eine bestimmte Reihenfolge gebracht werden.

39 / 40
Warum ist ästhetische Gestaltung wichtig?

A) Sie verbessert die Techniken des maschinellen Lernens.
B) Sie erleichtert die natürliche Sprachverarbeitung (NLP).
C) Sie konzentriert sich auf Modelle der Regressionsanalyse.
D) Sie macht die Datenvisualisierung benutzerfreundlicher.

A) Falsch. Maschinelles Lernen stützt sich nicht auf Datenvisualisierung, sondern basiert auf Datensätzen.
B) Falsch. NLP ist mit Sprache, Syntax und Semantik verbunden. Sie stützt sich nicht auf Datenvisualisierung.
C) Falsch. Die ästhetische Gestaltung verbessert alle Formen der Datenvisualisierung. Sie konzentriert sich nicht auf die Regressionsanalyse.
D) Richtig. Zu den Prinzipien der ästhetischen Gestaltung zählen sorgfältige Gestaltung sowie Farbkombinationen, die die Benutzerfreundlichkeit der Datenvisualisierung verbessern. (Literatur: A, Kapitel 11)

40 / 40
Welches Datenvisualisierungstool erinnert in Aussehen und Wirkung an Microsoft Excel?

A) Data Wrapper
B) Google Charts
C) Power BI
D) Tableau

A) Falsch. Data Wrapper hat eine eigene Schnittstelle.
B) Falsch. Google Charts ist mit Google-Produkten kompatible.
C) Richtig. Power BI ist mit den anderen Produkten von Microsoft kompatibel. (Literatur: A, Kapitel 12)
D) Falsch. Tableau hat eine eigene Schnittstelle.

Beurteilung

Die richtigen Antworten auf die Fragen in dieser Musterprüfung finden Sie in nachstehender Tabelle.

Frage	Antwort	Frage	Antwort
1	B	**21**	C
2	C	**22**	C
3	A	**23**	D
4	C	**24**	C
5	D	**25**	C
6	C	**26**	D
7	B	**27**	B
8	B	**28**	A
9	B	**29**	B
10	A	**30**	C
11	B	**31**	B
12	C	**32**	D
13	A	**33**	B
14	B	**34**	C
15	C	**35**	B
16	C	**36**	A
17	A	**37**	A
18	A	**38**	A
19	A	**39**	D
20	A	**40**	C

Kontakt EXIN

www.exin.com