Matrikelnummer: Name:

Data Science (46922)

Trial exam 2025

Notes:

- **Am Ende der Klausur finden Sie zur Unterstützung eine deutsche Übersetzung.**
- Please check the exam for completeness and enter your name and matriculation number on all sheets.
- Answer the questions in **German** or **English**.
- **Only** a non-programmable calculator and a handwritten A4 sheet of paper are allowed to be used as aids to complete the exam!
- If you carry out calculations with a calculator, provide at least 4 decimal places.
- Please make a corresponding note if you do not write your solution directly below or next to the task.
- If several solutions are given, the solution to be evaluated must be clearly marked. In addition to the correct solution to the problem, your solutions will also be assessed for comprehensible wording, sufficient documentation and clear writing. Also give intermediate steps of your solution.
- Please write clearly. Clearly cross out those parts of the texts you have written that should not be included in the assessment.

Good luck!

Aufgaben	Erreichbar	Erreicht
1	20	
2	18	
3	12	
4	15	
5	20	
6	15	
Summe	100	
Prozent	100 %	
Bonus	10 %	
Summe	110 %	

Exercise 1: **20 Punkte**

Provide an answer to the following questions.

(a) What is the CRISP-DM process model and which steps are included? (4)

(b) What is the difference between the JSON and CSV data format? (4)

(c) What is the difference between primary and secondary data? (4)

(d) What is the concept of a document-oriented databases? (4)

(e) What is a hash-function and why is it useful for data protection? (4)

Exercise 2: **18 Punkte**

Commuter traffic is measured on a bicycle route at a turn-off. The following absolute frequencies are measured, depending on the time and direction:

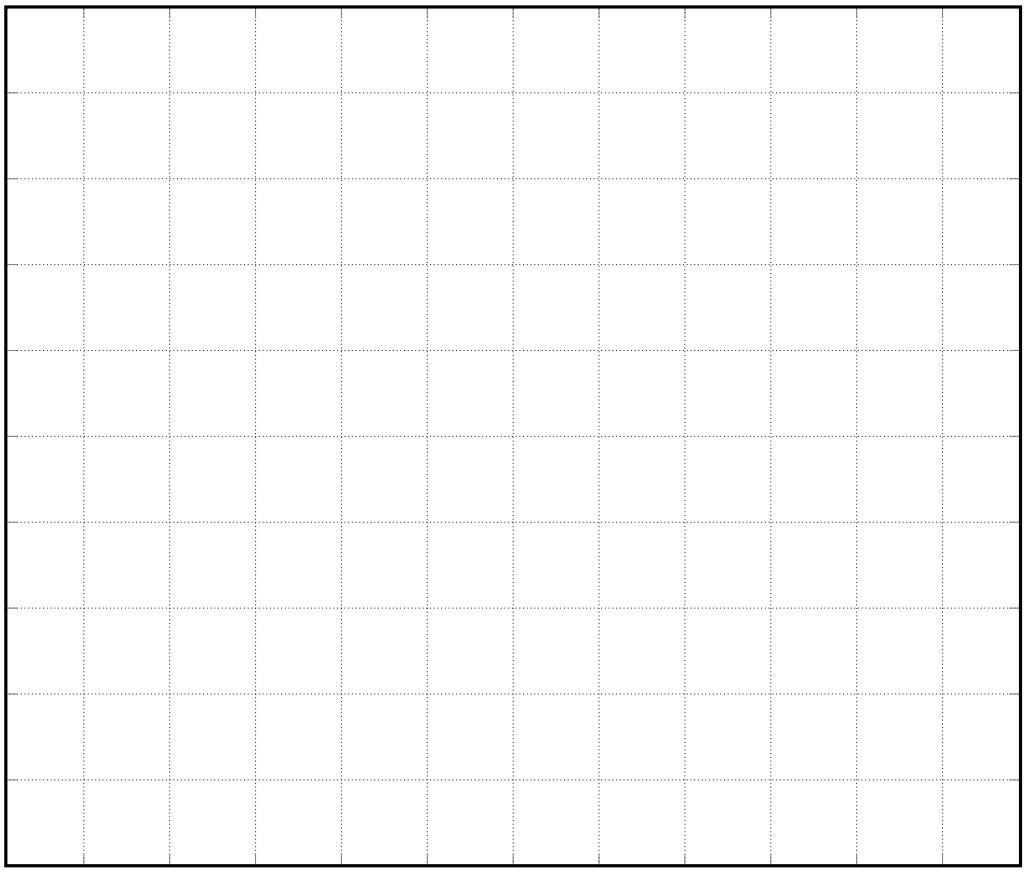
	morning	**noon**	**evening**	**night**
city	200	135	50	10
surroundings east	25	10	150	5
surroundings south	25	5	50	5

(a) Compute the marginal frequencies. How large is the proportion of rides, independent of the daytime, in direction city and the surroundings? (5)

(b) For every direction: What is the modal value? (2)

(c) Among the rides in direction city: Which proportion have rides in the morning and the evening? (4)

(d) Create a plot showing the conditional relative frequencies of the rides in direction surrounding east for the different times. (4)



(e) Compare the rides in the morning in direction city and in the evening in direction surroundings east together with surroundings south. What can you observe? How can this be explained? (3)

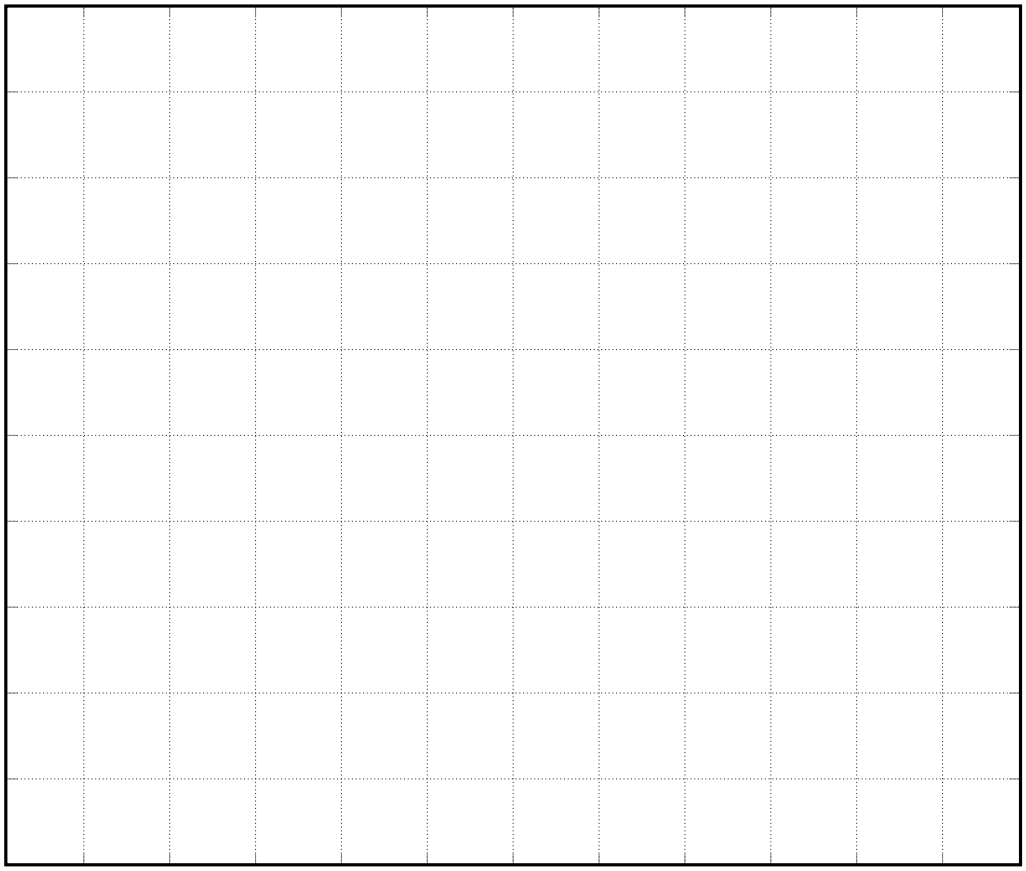
Exercise 3: **12 Punkte**

A team of scientists of the FH Dortmund measures the time, students need to get from EFS 42 to the cafetaria kostbar. They observed, that the speed, the students walk, can be described by a linear probability density function in the range of $0km/h - 5km/h$, where the probability of $0km/h$ is 0 and for $5km/h$ is the highest.

> **Hint:** A linear density function is defined as $f(x) = a \cdot x + b$, where a and b are real values. The corresponding antiderivative is given by $F(x) = \frac{a}{2}x^2 + b \cdot x$.

(a) Derive the probability density from the given information. (5)

(b) Compute the probability distribution function and sketch the function in a proper way. (4)



(c) What is the probability that a person walks between $2km/h$ and $4km/h$? (3)

Exercise 4: **15 Punkte**

The relationship between income and happiness was examined in a survey. Four respondents were asked about their annual net income (in 1000 euros) and their happiness (score from 1 to 10) was determined:

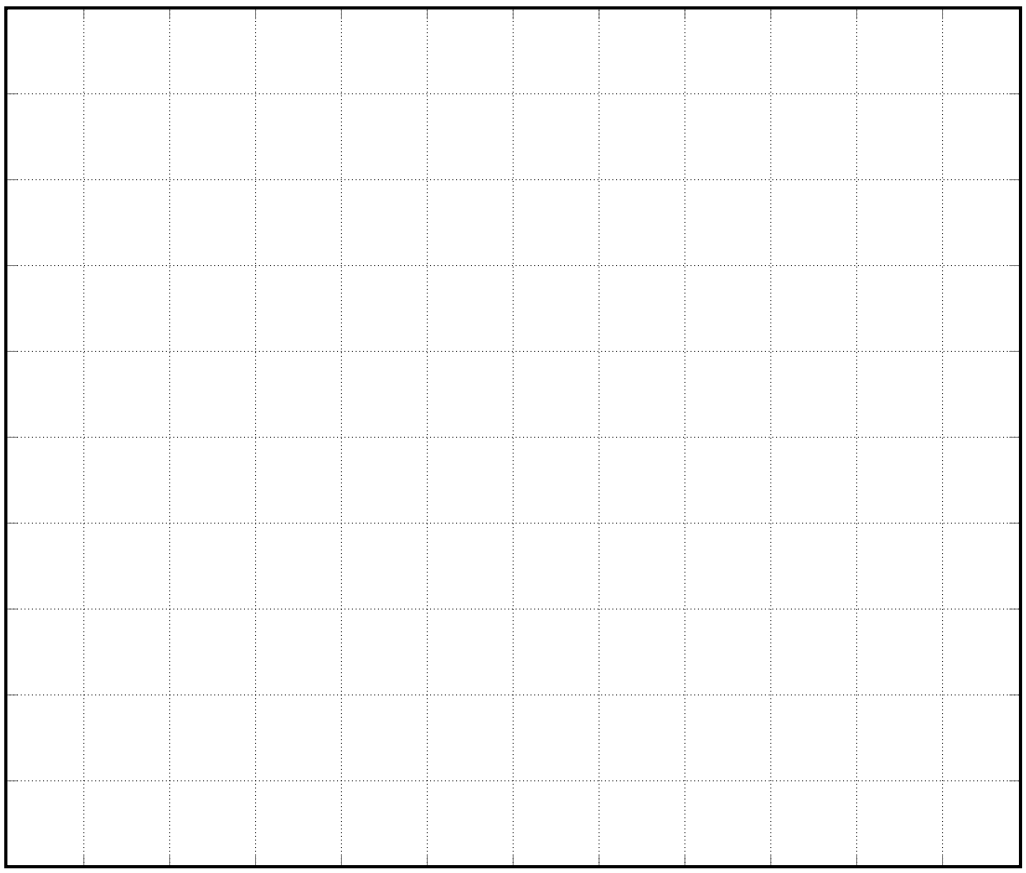
income	20	20	15	25
happiness	6	8	5	9

The goal is to predict the happiness for a given income.

(a) Compute and interpret the Pearson correlation coefficient. (4)

(b) Perform a simple linear regression, i.e. compute the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Also provide and interpret the resulting regression line. (6)

(c) Draw a scatter plot of the data and add the regression line (3)



(d) Compute the coefficient of determination for the given regression. (2)

Exercise 5: **20 Punkte**

A research group is investigating the success of books and has collected the following data:

length	**genre**	**hardcover**	**success**
short	thriller	yes	high
short	romance	yes	low
long	romance	no	high
medium	romance	no	high
medium	thriller	yes	low
long	thriller	no	high

The researcher want to predict the success of books based on the given variables.

(a) Train a decision tree classifier, based on the method given in lecture (gain of order). Draw the final decision tree. (15)

(b) Classify the following data concerning their success. Describe the path taken in the decision tree! (5)

length	**genre**	**hardcover**
short	romance	yes
medium	romance	no
medium	thriller	yes

Exercise 6: **15 Punkte**

In recent years, it has been shown that the points achieved in the "Data Science" exam follow a normal distribution with a mean of 65 and a standard deviation of 10. The following points were achieved in the past exam:

90 45 65 70 85 92 30 80

Hint: The values of the z and t (for different n) distribution are given in the following table.

	0.9	**0.95**	**0.99**
z	1.282	1.645	2.326
t(n=7)	1.415	1.895	2.998
t(n=8)	1.397	1.860	2.896
t(n=9)	1.383	1.833	2.821

(a) Compute the average value of the points. (2)

(b) Can you confirm the following statement with a significance of 0.1? (13)

"The results this year are better than in previous years"

Carry out a suitable statistical test for this.

Deutsche Übersetzung

Wichtig: Bitte beachten Sie, dass die deutschen Übersetzungen zur Erleichterung dienen; die maßgeblichen Aufgabenstellungen sind jedoch die englischen Versionen!

Hinweise

- Bitte überprüfen Sie die Klausur auf Vollständigkeit und schreiben Sie Name und Matrikelnummer auf alle Seiten.
- Als Hilfsmittel sind nur ein nicht programmierbarer Taschenrechner und ein handbeschriebenes Din A4 Blatt erlaubt.
- Wenn Sie Berechnungen mit dem Taschenrechner machen, geben Sie bitte mindestens 4 Nachkommastellen an.
- Machen Sie einen Hinweis, wenn Sie eine Aufgabe nicht direkt unter oder neben der Aufgabenstellung bearbeiten.
- Wenn mehrere Lösungen gegeben sind, muss die zu bewertende Lösung eindeutig gekennzeichnet sein. Bei Ihren Lösungen werden außer der korrekten Problemlösung auch die verständliche Formulierung, ausreichende Dokumentation und eine übersichtliche Schreibweise bewertet.
- Schreiben Sie deutlich. Streichen Sie Sachen, die nicht bewertet werden sollen deutlich durch.

Aufgabe 1: **20 Punkte**

Geben Sie eine Antwort auf die folgenden Fragen.

(a) Was ist der CRISP-DM Prozess und welche Schritte beinhaltet dieser? (4)

(b) Was ist der Unterschied zwischen dem JSON und CSV Datenformat? (4)

(c) Was ist der Unterschied zwischen Primären und Sekundären Daten? (4)

(d) Was ist das Konzept von Dokumentenorientierten Datenbanken? (4)

(e) Was ist eine Hash-funktion und warum ist diese für den Datenschutz hilfreich? (4)

Aufgabe 2: **18 Punkte**

Pendlerverkehr wird auf einer Fahrradstrecke an einer Abzweigung gemessen. Die folgenden absoluten Häufigkeiten werden gemessen, abhängig von der Zeit und Fahrtrichtung:

	morning	**noon**	**evening**	**night**
city	200	135	50	10
surroundings east	25	10	150	5
surroundings south	25	5	50	5

(a) Berechne die Randhäufigkeiten. Wie groß ist der Anteil an Fahrten, unabhängig von der Tageszeit, in Richtung Innenstadt und in die Umgebung? (5)

(b) Für jede Richtung: Was ist der Modalwert? (2)

(c) Unter den Fahrten in Richtung Innenstadt: Welchen Anteil haben Fahrten am morgen und am Abend? (4)

(d) Erstelle einen Plot, der die bedingten relativen Häufigkeiten von Fahrten in die Richtung Umgebung Ost für die verschiedenen Zeiträume zeigt. (4)

(e) Vergleiche die Fahrten am Morgen in Richtung Innenstadt und am Abend in Richtung Umgebung Ost zusammen mit Umgebung Süd. Was können Sie Beobachten? Wie lässt sich dies erklären? (3)

Aufgabe 3: **12 Punkte**

Ein Team von Wissenschaftlern der FH Dortmund misst die Zeit, welche Studierende benötigen, um von der EFS 42 zur Mensa kostbar zu kommen. Sie beobachten, dass die Geschwindigkeit, welche die Studierenden laufen, durch eine lineare Wahrscheinlichkeitsdichte im Bereich von $0km/h - 5km/h$, wobei die Wahrscheinlichkeit von $0km/h$ 0 ist und die für $5km/h$ am höchsten ist, beschrieben werden kann.

> **Hinweis:** Eine lineare Dichtefunktion ist definiert als $f(x) = a \cdot x + b$, wobei a und b reelle Zahlen sind. Die Stamfunktion ist entsprechend $F(x) = \frac{a}{2}x^2 + b \cdot x$.

(a) Bestimme die Wahrscheinlichkeitsdichte basierend auf diesen Informationen. (5)

(b) Berechne die Wahrscheinlichkeitsverteilungsfunktion und skizziere diese. (4)

(c) Was ist die Wahrscheinlichkeit das eine Person zwischen $2km/h$ und $4km/h$ läuft? (3)

Aufgabe 4: **15 Punkte**

Der Zusammenhang zwischen Einkommen und Glück wurde in einer Umfrage untersucht. Vier Personen wurden befragt und Ihr jährliches Einkommen (in 1000 Euro) aufgenommen und Ihr Glück (Score zwischen 1 und 10) wurde bestimmt.

income	20	20	15	25
hapiness	6	8	5	9

Das Ziel ist es das Glück anhand des Einkommens vorherzusagen.

(a) Berechne und Interpretiere den Pearson Korrelationskoeffizient. (4)

(b) Führe eine lineare Einfachregression durch, d.h. berechne die Parameterschätzer $\hat{\beta}_0$ und $\hat{\beta}_1$. Gib die resultierende Regressionslinie an und Interpretiere diese. (6)

(c) Zeichne einen Scatter Plot von den Daten und füge die Regressionslinie hinzu. (3)

(d) Berechne das Bestimmtheitsmaß der gegebenen Regression. (2)

Aufgabe 5: **20 Punkte**

Eine Gruppe Wissenschaftler untersucht den Erfolg von Büchern und hat folgende Daten gesammelt:

length	genre	hardcover	success
short	thriller	yes	high
short	romance	yes	low
long	romance	no	high
medium	romance	no	high
medium	thriller	yes	low
long	thriller	no	high

Die Wissenschaftler wollen den Erfolg der Bücher anhand der gegebenen Variablen vorhersagen.

(a) Trainiere einen Entscheidungsbaum, basierend auf der Methode gegeben in der Vorlesung (gain of order). Zeichne der finalen Entscheidungsbaum. (15)

(b) Klassifiziere die folgenden Datenpunkte bezüglich ihrem Erfolg. Beschreibe den Pfad der im Entscheidungsbaum genommen wurde. (5)

length	genre	hardcover
short	romance	yes
medium	romance	no
medium	thriller	yes

Aufgabe 6: **15 Punkte**

In den vergangenen Jahren hat sich gezeigt das die Punkte, welche Studierende in der "Data Science" Klausur erreicht haben, einer Normalverteilung mit Mittelwert 65 und einer Standardabweichung von 10 folgen. Die folgenden Punkte wurden in der letzten Klausur erreicht:

90 45 65 70 85 92 30 80

Hint: Die Werte der z und t (für verschiedene n) Verteilung sind in der folgenden Tabelle gegeben.

	0.9	0.95	0.99
z	1.282	1.645	2.326
t(n=7)	1.415	1.895	2.998
t(n=8)	1.397	1.860	2.896
t(n=9)	1.383	1.833	2.821

(a) Berechne den Mittelwert dieser Datenpunkte (2)

(b) Können Sie die folgende Aussage mit einer Signifikanz von 0.1 zeigen? (13)

"Die Ergebnisse in diesem Jahr sind besser als in den Jahren zuvor"

Führen Sie hierfür einen passenden Statistischen Test durch.