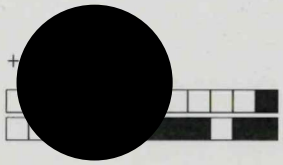


Processing notices concerning the whole exam

- **Corona measures:**
  - Wearing a mask is mandatory before, during, and after the exam. (Surgical masks are the minimum.)
  - Exception: If the distance to all seat neighbors is at least 1.5m, the mask may be removed during the test.
- **Exam details:**
  - The duration of the exam is 90 minutes.
  - Check your examination copy for completeness (in total 15 pages).
- **Utilities:**
  - You may bring any number of writing utensils (pens, rulers, erasers).
  - You should bring at least one correction roller.
  - You can prepare a single hand-written sheet (paper format A4, only front page), any printing is forbidden. The sheet's contents are not subject of any restrictions.
  - Any other documents or utilities are forbidden.
- **Workplace:**
  - Bags and wardrobe must be stored near the workplaces.
  - Electronic devices are to be switched off and stored in their pockets before the start of the exam. Exception: Corona-Warn-App.
  - Have your photo ID ready at your seat (identity check).
- **Questions:**
  - Use only document-proof pens in shades of dark blue or black for your answers. We reserve the right not to consider answers in pencil for correction.
  - Read the task carefully and pay attention to negations like not or whether you should identify correct or wrong answers.
  - **Multiple choice questions**:
    - Fill in the boxes completely.
    - Use only correction rollers for corrections. Make sure that you cover the box completely with correction tape. If you want the box to count again, paint over the correction tape at the position of the original box.
    - The use of correction fluid and ink eraser is not permitted for technical reasons.
  - **Free-response questions**:
    - Try to answer all questions unambiguously. Answers that are obviously incorrect or needless can cause a deduction of points.
    - The answers to the questions of this examination must be given in English. Answers in other languages will not be considered.
    - Do not answer outside of the template boxes! If you need more space, you can use the back pages of your exam copy. However, take care that your pen's colour does not shine through to the front. In the original template box refer to the back page where your solution can be found.
    - Clearly cross out any answers that are not to be evaluated.
- **Other notices:**
  - Do not remove the retaining clip of your examination copy and do not manipulate the codes at the upper margin, the grading fields and the calibration symbols in every corner. Each manipulation will be graded as an attempt to deceive and will lead to failing the exam.
  - We collect your examination copy and your prepared cheat sheet at the end of the exam. You are not allowed to take any sheet with you! We evaluate only answers on the original examination copy!
  - If you have to cancel the exam for health reasons, inform us immediately. You will then receive a form and must obtain a medical certificate from a medical officer appointed by the university.

The entire KDDmUe team wishes you lots of success!

[redacted]

Introduction (Processing time approx. 6 min)

Frage 1
For the following statements, mark whether they are **True** or **False** with regard to **data mining**.
Mark one applicable answer per statement (3 completely filled boxes in total).

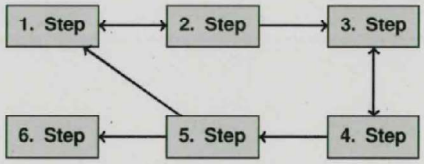
☐ **True** ☒ **False** The input data for data mining must always be stored in a relational database.

6/6

☒ **True** ☐ **False** Knowledge discovery and information harvesting are closely related to data mining.

☒ **True** ☐ **False** The extraction of interesting patterns from huge amounts of data is called data mining.

Frage 2
Given is the outline of the **CRoss-Industry Standard Process for Data Mining** (CRISP-DM):



For each of the following steps, mark **which step** (1st-6th) of CRISP-DM it is.
Mark one applicable answer per step (6 completely filled boxes in total).

☐1 ☐2 ☒3 ☐4 ☐5 ☐6 **Step:** Data preparation

☐1 ☒2 ☐3 ☐4 ☐5 ☐6 **Step:** Data understanding

☐1 ☐2 ☐3 ☐4 ☐5 ☒6 **Step:** Deployment

6/6

☐1 ☐2 ☐3 ☐4 ☒5 ☐6 **Step:** Evaluation

☒1 ☐2 ☐3 ☐4 ☐5 ☐6 **Step:** Business understanding

☐1 ☐2 ☐3 ☒4 ☐5 ☐6 **Step:** Modelling

[redacted]

Data (Processing time approx. 7.5 min)

Frage 3

One of the first taks in data analysis is to get to know the data at hand. Which of the following statements regarding **statistical descriptors and plots** are **False**?
Mark all applicable answers (1 - n completely filled boxes).

- [x] A distribution is positively skewed when the mean is smaller than the mode.
- [x] The median of interval grouped data lies in the group which exceeds 50% of the relative frequency.
- [ ] A bimodal dataset is a dataset that has two frequently occuring values.
- [x] A bar chart is used to asses the probability distribution of a dataset's attribute.
- [x] The Inter-Quartile-Range is affected by extreme values.
- [ ] A scatter plot is used to asses patterns or the type of relationship between two attributes.
- [ ] Standard deviation is a measure of dispersion signaling how close values are to the mean.

6/8

Frage 4

For the following statements, mark whether they are **True** or **False** with regard to the **types of attributes**.
Mark one applicable answer per statement (7 completely filled boxes in total).

● **True** ☒ **False**	Continuous attributes have countably finite elements.
☒ **True** ☐ **False**	Interval-scaled attributes are ordered.
☐ **True** ☒ **False**	Continuous attributes have a true zero point.
☐ **True** ☒ **False**	Measurements can be divided into qualitative, quantitative, and categorical attributes.
☒ **True** ☐ **False**	Symmetric binary attributes are a specialization of discrete attributes.
☒ **True** ☐ **False**	Nominal attributes are unordered.
● **True** ☒ **False**	The terms *types of measurements* and *types of attributes* are synonyms.

3/7

Preprocessing (Processing time approx. 16 min)

Frage 5
Which of the following **are among** the measures of data quality?
Mark all applicable answers (1 - n completely filled boxes).

2/4

- [x] Interpretability
- [ ] Velocity
- [x] Consistency
- [x] Timeliness
- [x] Dimensionality
- [x] Accuracy

The **next two** questions refer to the following dataset about exam participants and their grades. The dataset contains different types of dirty data. You may assume that all types of dirty data **are clearly identifiable**.

Name	Course of Study	Points	Grade
John Doe	Informatics, B.Sc.	91	4.0
John Doe	Bachelor in Informatics	178	-
John Doe	Data Science, B. Sc.	-10	1.3
John Doe	Data Science, M.Sc.	178	A

Frage 6
Which of the attributes contain **errors**?
Mark all applicable answers (1 - n completely filled boxes).

0/2

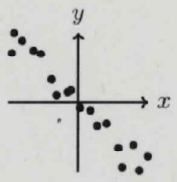
- [x] Name
- [ ] Course of Study
- [x] Points
- [ ] Grade

Frage 7
Which of the attributes contain **inconsistencies**?
Mark all applicable answers (1 - n completely filled boxes).

4/4

- [ ] Name
- [x] Course of Study
- [ ] Points
- [x] Grade

Frage 8
What kind of correlation is indicated in the **diagram on the right**?
Mark one applicable answer (1 completely filled box).



3/3

- [x] Negative correlation
- [ ] Positive correlation
- [ ] Correlation, but it can not be said if positive or negative
- [ ] Uncorrelated/no correlation

Frage 9
What type of correlation is indicated when the **chi-squared test** gives a value of **236**?
Mark one applicable answer (1 completely filled box).

3/3

- [ ] Negative correlation
- [ ] Positive correlation
- [x] Correlation, but it can not be said if positive or negative
- [ ] Uncorrelated/no correlation

[redacted]

Frage 10
Given is a **sorted list** of account balances:

$$[-97000, -2132, -421, 98, 200, 2306, 3000]$$

To which value is **200** normalized when applying **normalization by decimal scaling** to this list?
Mark one applicable answer (1 completely filled box).

4/4

☐ 0.369	☐ 0.972	☐ 0.001	☐ 0.020
☐ 0.097	☒ 0.002	☐ 0.003	☐ 0.100

Frage 11
Given is a **mystery normalization** function:

```
1 def mystery_normalization(df):
2   return (df - df.mean()) / df.std()
```

Which of the following normalization functions **is implemented** in this function?
Mark one applicable answer (1 completely filled box).

3/3

☐ Kulczynski normalization	☒ Z-score normalization
☐ Normalization by decimal scaling	☐ Min-Max normalization

Frage 12
Given is a **sorted list** of temperature values:

$$[-3, -2, 1, 23]$$

Which of the values end up in **Bin 1** (lower values) and which in **Bin 2** (higher values) when **equal-width partitioning** with **two bins** is performed on the list?
Mark one applicable answer per value (4 completely filled boxes in total).

4/4

☒ **Bin 1** ☐ **Bin 2** -3		☒ **Bin 1** ☐ **Bin 2** 1	
☒ **Bin 1** ☐ **Bin 2** -2		☐ **Bin 1** ☒ **Bin 2** 23	

Frage 13
For each step, mark **which step** (1st-5th) within a **Principal Component Analysis** (PCA) it is. The standardization is considered as part of the PCA in this case.
Mark one applicable answer per step (5 completely filled boxes in total).

5/5

☐ 1	☐ 2	☒ 3	☐ 4	☐ 5	**Step:** Compute the eigenvectors and eigenvalues
☐ 1	☐ 2	☐ 3	☐ 4	☒ 5	**Step:** Transform the data
☐ 1	☐ 2	☐ 3	☒ 4	☐ 5	**Step:** Construct a feature matrix
☒ 1	☐ 2	☐ 3	☐ 4	☐ 5	**Step:** Standardize the data
☐ 1	☒ 2	☐ 3	☐ 4	☐ 5	**Step:** Compute a covariance matrix

[redacted]

OLAP (Processing time approx. 4.5 min)

Frage 14

In his definition of a data warehouse, William H. Inmon attributed certain properties to the data warehouse. Which of the following **are among** these properties?

Mark all applicable answers (1 - n completely filled boxes).

2/3

- [x] Subject-oriented
- [ ] Multi-dimensional
- [ ] Consolidated
- [x] Heterogeneous
- [x] Integrated
- [ ] Volatile
- [x] Time-variant

Frage 15

For the following statements, mark whether they are **True** or **False** with regard to the **conceptual modelling of a data warehouse**.

Mark one applicable answer per statement (6 completely filled boxes in total).

4/6

[x] **True** [ ] **False**	Fact tables in a fact constellation can share measurements.
[ ] **True** [x] **False**	Fact constellations, also called galaxy schema, share dimensions of multiple snowflake schemas but dimensions from multiple star schemas.
[ ] **True** [x] **False**	Star schemas have a separate dimension for time and measurements. (not)
[ ] **True** [x] **False**	Snowflake schema has dimensions that strictly follow a third normal form.
[ ] **True** [x] **False**	Fact constellations are not subject-oriented because they contain multiple subjects in the form of fact tables.
[x] **True** [ ] **False**	Virtual Warehouse is a set of views over operational databases.

Frequent Patterns (Processing time approx. 15 min)

Frage 16
Given is a **complete** set of itemsets for a dataset and their respective occurence counts:

$\{Knife\}$	2
$\{Bowl\}$	3
$\{Pan\}$	2
$\{Gloves\}$	1

$\{Knife, Bowl\}$	2
$\{Knife, Pan\}$	1
$\{Bowl, Pan\}$	2

$\{Knife, Bowl, Pan\}$	1

For the following statements, mark whether they are **True** or **False** with regard to the above **set of itemsets**.
Mark one applicable answer per statement (3 completely filled boxes in total)

☒ **True** ☐ **False** Given a minimum support count of 2, $\{Bowl, Pan\}$ would be a closed itemset.

6/6

☒ **True** ☐ **False** Given a minimum support count of 2, $\{Knife, Bowl\}$ would be a max-itemset.

☐ **True** ☒ **False** Given a minimum support count of 2, $\{Gloves\}$ would be a frequent itemset.

Frage 17
Given are **all** frequent 2-itemsets for a dataset:

$$\{\{Football, Shoes\}, \{Football, Jersey\}, \{Whistle, Jersey\}, \{Jersey, Shoes\}\}$$

Based on these frequent 2-itemsets, which itemsets would be part of **the result** of the next candidate generation of **a Priori**?
Mark all applicable answers (1 - n completely filled boxes).

4/4

- ☐ $\{Football, Whistle, Jersey\}$
- ☒ $\{Football, Shoes, Jersey\}$
- ☐ $\{Whistle, Shoes, Jersey\}$
- ☐ $\{Football, Shoes, Gloves\}$
- ☐ $\{Football, Shoes, Whistle\}$
- ☐ $\{Shoes, Whistle\}$

Frage 18
For an association rule $Schnitzel \rightarrow Beer$ the following values were calculated via the **Kulczynski Measure** ($Kulc$) and the **Imbalance Ratio** (IR):

$$Kulc = 0.01$$
$$IR = 0.99$$

For the following statements, mark whether they are **True** or **False** with regard to the above **values**.
Mark one applicable answer per statement (2 completely filled boxes in total)

☒ **True** ☐ **False** Based on the IR value it can be said that the association rule is unbalanced.

0/4

☒ **True** ☒ **False** Based on the $Kulc$ value it can be said that $Schnitzel$ and $Beer$ are associated.

Frage 19
Given is the **transactional dataset shown on the right**.
Using the FPGrowth approach, create the initial FP-Tree for this dataset. Use a minimal support count of 2.

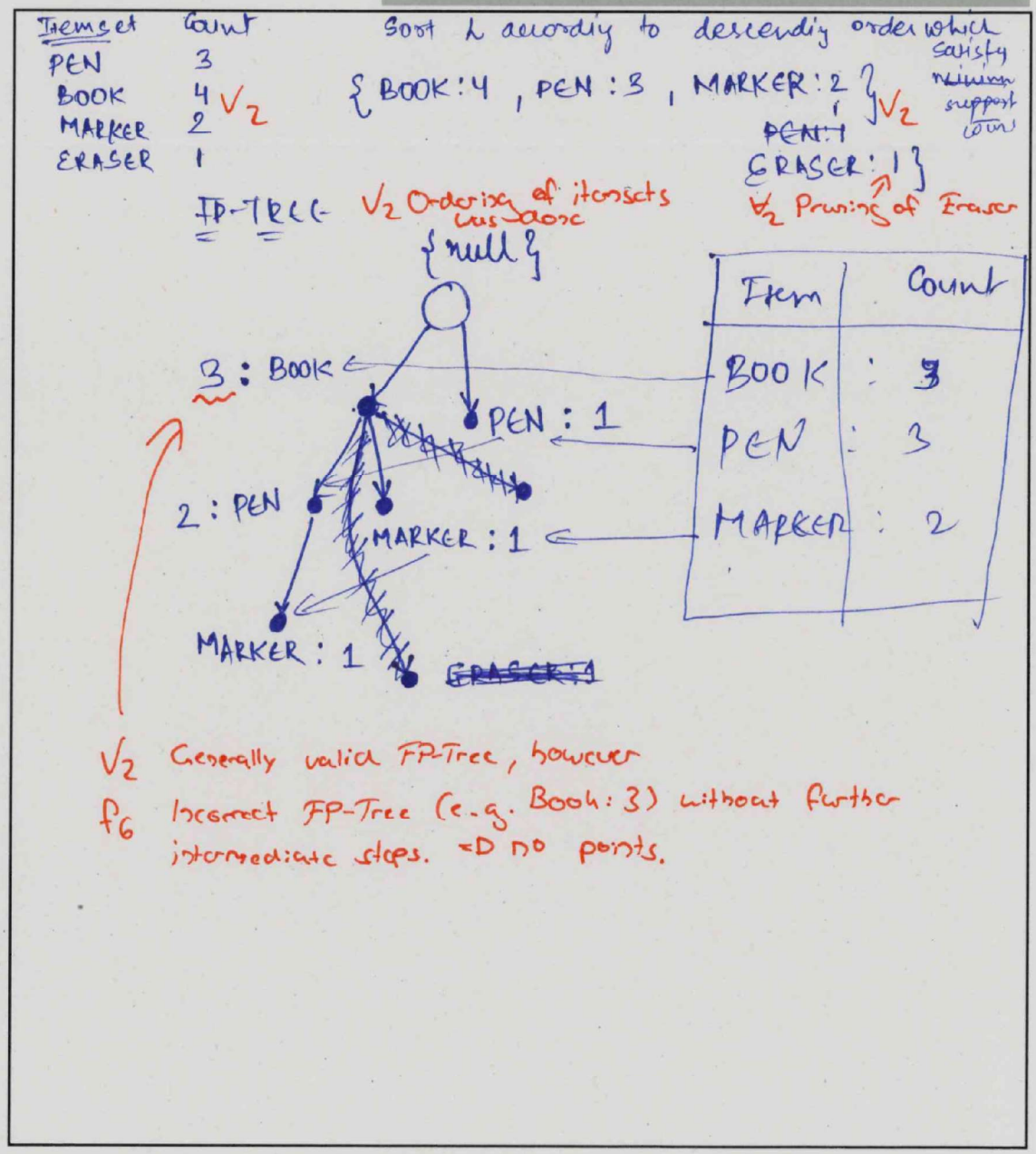
Note: Intermediate steps do not have to be written down, but if the final result is wrong, they are used to award partial points. We therefore recommend to write down at least some intermediate steps.

TID	Items bought
1	Pen, Book, Marker
2	Book, Eraser
3	Marker, Book
4	Pen
5	Book, Pen

Only for grading (do not fill in!):

☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☒8 ☐9
☐10 ☐11 ☐12 ☐13 ☐14 ☐15 ☐16

8/16



V2 Generally valid FP-Tree, however
f6 Incorrect FP-Tree (e.g. Book: 3) without further intermediate steps. =D no points.

[redacted]

Classification (Processing time approx. 21 min)

Frage 20

We have discussed three methods for **selecting attributes** to create a decision tree. Assign the following statements to the **corresponding method** or mark **None** if they cannot be matched to any of the methods.
Mark one applicable answer per statement (6 completely filled boxes in total).

☒ Information gain	☐ Gain ratio	☐ Gini index	● None	Guarantees a simple tree.
☐ Information gain	☐ Gain ratio	☒ Gini Index	☐ None	Select attribute as splitting attribute with the lowest function value.
☐ Information gain	☒ Gain ratio	☐ Gini index	☐ None	Prefers unbalanced splits.
☐ Information gain	● Gain ratio	☒ Gini Index	☐ None	Enforces a binary tree.
● Information gain	☐ Gain ratio	● Gini index	☒ None	Biased towards discrete attributes.
☒ Information gain	☐ Gain ratio	☐ Gini index	☐ None	Used by ID3.

0/6

Frage 21

Given is a function that belongs to an attribute selection method:

```
1  def attribute_selection_helper(dataset: pd.DataFrame, partition_attribute: str) -> float:
2      weights = dataset[partition_attribute].value_counts() / dataset.shape[0]
3      return sum([weight * log(weight, 2) for _, weight in weights.items()]) * -1
4
```

In which of the following **attribute selection method** is this function most likely to **be used**?
Mark all applicable answers (1 - n completely filled boxes).

☒ Gain ratio ☒ Information gain ☐ Gini index

2/2

Frage 22

For the following statements, mark whether they are **True** or **False** with regard to **ensemble methods in general**.
Mark one applicable answer per statement (7 completely filled boxes in total).

☒ True	☐ False	Bagging seeks to reduce variance by sampling a dataset with replacement.
☐ True	☒ False	Bagging requires models to be trained sequentially to determine weights for each training tuple.
☐ True	☒ False	The .632 bootstrap method assigns a weight to each tuple.
● True	☒ False	AdaBoost is more robust to errors and outliers than a random forest if used on a suffiently large dataset.
☐ True	☒ False	Random forests have a problem of overfitting when a large number of trees are combined.
☒ True	☐ False	Random forests use bagging and random attribute selection.

4/6

Frage 23

A new Pokemon has been discovered and your abilities to determine the legendary status are needed. The Pokemon in question is extremely shy, but a Data Engineer gathered the following properties:

$$X = \{\text{'type': 'psychic', 'defense': 'medium', 'speed': 'high'}\}.$$

Unfortunately, the psychic abilities of this Pokemon is quite pronounced rendering your computer unusable. This leaves you with only a small test dataset - a test dataset that you know by heart. This dataset is displayed on the right hand side. Therefore, the only reasonable method to determine the legendary status is to compute **naive Bayes** manually.

Calculate the needed values to determine the legendary status of this new Pokemon.

number	name	defense	speed	legendary
1	Bulbasaur	low	low	no
4	Charmander	low	medium	no
7	Squirtle	low	low	no
97	Hypno	low	medium	no
142	Aerodactyl	medium	high	no
144	Articuno	medium	medium	yes
145	Zapdos	medium	medium	yes
146	Moltres	medium	medium	yes
150	Mewtwo	low	high	yes
151	Mew	medium	medium	yes

Note: In case of zero probabilities, do *not* use Laplacian correction. *Fractions* as result are sufficient. No need to convert them to rational numbers.

Only for grading (do not fill in!):

☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9
☐10 ☐11 ☒12

12/12

$P(C_i = \text{yes}) = \frac{5}{10}$ ✓ $\quad P(C_i = \text{no}) = \frac{5}{10}$ ✓

~~$P(\text{type} = \text{"psychic"} \mid \text{legendary})$~~

$P(\text{defense} = \text{medium} \mid \text{legendary} = \text{yes}) = \frac{4}{5}$ ✓

$P(\text{defense} = \text{medium} \mid \text{legendary} = \text{no}) = \frac{1}{5}$ ✓

$P(\text{speed} = \text{high} \mid \text{legendary} = \text{yes}) = \frac{1}{5}$ ✓

$P(\text{speed} = \text{high} \mid \text{legendary} = \text{no}) = \frac{1}{5}$

$P(X_i \mid C) = \prod P(x_i \mid C)$

$P(X_i \mid C_i = \text{yes}) = \frac{4}{5} * \frac{1}{5} = \frac{4}{25}$ ✓

$P(X_i \mid C_i = \text{no}) = \frac{1}{5} * \frac{1}{5} = \frac{1}{25}$ ✓

Page No.

$P(C_i \mid X_i)_{\text{yes}} = P(X_i \mid C_i = \text{yes}) \cdot P(C_i = \text{yes})$ ½

$= 4/25 * \frac{5}{10} = 2 = \frac{4}{50}$ ✓ answer on backside

answer on back

Frage 24

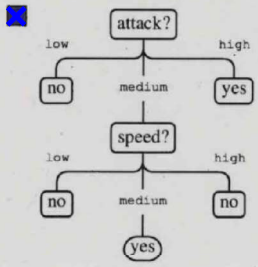
Given are four figures, where each figure shall depict a decision tree after training on a different Pokemon dataset. This different Pokemon dataset contains an additional attribute called "attack" with only two possible values, namely `low` and `high`. The following decision trees are not pruned or optimized. *Information gain* has been used as the attribute selection method.

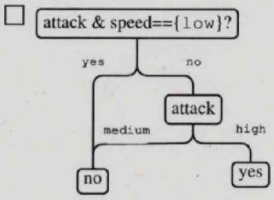
medium

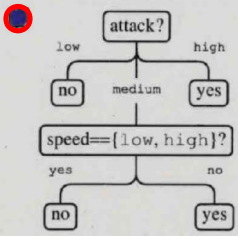
Which figures show **valid decision trees**?

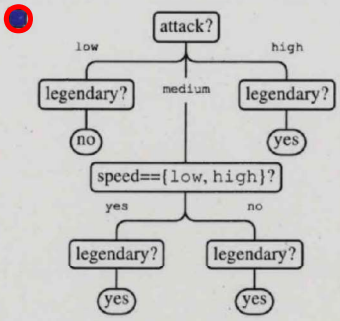
Mark all applicable answers (1 - n completely filled boxes).

0/4









Frage 25

One of the problems of a decision tree is **overfitting** which can be handled with tree pruning. For the following statements, mark whether they are **True** or **False** with regard to **decision tree pruning**.

Mark one applicable answer per statement (4 completely filled boxes in total).

0/8

☒ True ☐ False	Branches may reflect noise or outliers. Therefore, these branches must be pruned.
☒ True ☒ False	Prepruning is a technique to halt tree construction early if the accuracy falls below 50%.
☒ True ☒ False	Pruned trees are typically smaller, less complex, easier to comprehend, faster and better at classifying unseen data.
☐ True ☒ False	Given a fully grown tree, randomly removing branches or subtrees increases the overall accuracy.

Frage 26

Imagine that the cells of the confusion matrix on the right hand side are filled with numbers after evaluating a classifier that has been trained with the Pokemon test dataset. This classifier predicts the legendary status of a new Pokemon (legendary, or not legendary). Match the cells (1 - 4) to the correct statements with regard to a Pokémon test dataset.

		True Class	
		$\neg C$	C
Predicted Class	$\neg C$	**Cell 1**	**Cell 2**
	C	**Cell 3**	**Cell 4**

Mark one applicable answer per cell (4 completely filled boxes in total).

4/4

☒ 1	☐ 2	☐ 3	☐ 4	Number of Pokemons that are not legendary correctly predicted as not legendary.
☐ 1	☐ 2	☐ 3	☒ 4	Number of Pokemons that are legendary correctly classified as legendary.
☐ 1	☒ 2	☐ 3	☐ 4	Number of Pokemons that are legendary incorrectly classified as not legendary.
☐ 1	☐ 2	☒ 3	☐ 4	Number of Pokemons that are not legendary incorrectly classified as legendary.

[redacted]

Clustering (Processing time approx. 14 min)

Frage 27

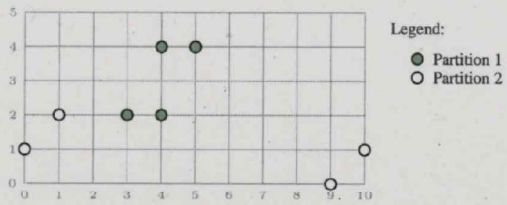
For the following statements, mark whether they are **True** or **False** with regard to **clustering in general**.
Mark one applicable answer per statement (4 completely filled boxes in total).

8/8

- ☐ **True** ☒ **False** Clustering methods are part of the supervised learning methods.
- ☒ **True** ☐ **False** A good clustering method results in high intraclass similarity.
- ☒ **True** ☐ **False** The quality of a clustering method depends partly on the similarity measure used.
- ☐ **True** ☒ **False** Clustering methods can process a maximum of two dimensions in one pass.

Frage 28

Given is the **iterim status** of a **k-means clustering** ($k = 2$) run-through:



Which points end up in **Partition 1** and which in **Partition 2** after the **next reassignment** using the **euclidean distance**?
Mark one applicable answer per point (8 completely filled boxes in total).

8/8

☒ **Partition 1**	☐ **Partition 2**	(0,1)	☒ **Partition 1**	☐ **Partition 2**	(4,4)
☒ **Partition 1**	☐ **Partition 2**	(1,2)	☒ **Partition 1**	☐ **Partition 2**	(5,4)
☒ **Partition 1**	☐ **Partition 2**	(3,2)	☐ **Partition 1**	☒ **Partition 2**	(9,0)
☒ **Partition 1**	☐ **Partition 2**	(4,2)	☐ **Partition 1**	☒ **Partition 2**	(10,0)

Frage 29

Given is a **mystery helper** function:

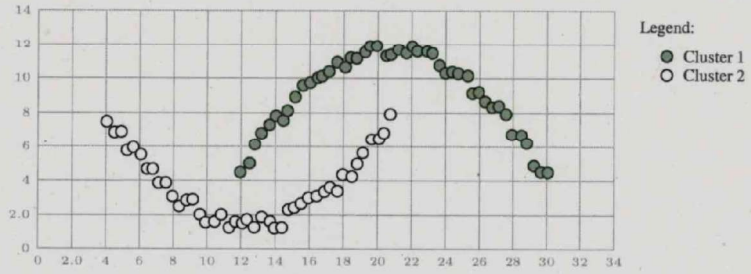
```
1  def mystery_helper(point, df, eps):
2    return df[
3      df.apply(
4        lambda a: mystery_distance(a, point) <= eps,
5        axis=1,
6      )
7    ]
```

In which of the following **clustering** algorithms is this function most likely to **be used**?
Mark one applicable answer (1 completely filled box).

4/4

- ☐ DIANA
- ☐ BIRCH
- ☐ STING
- ☒ DBSCAN

Frage 30
Given is the following **clustering result**:



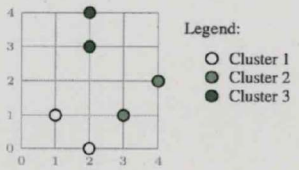
Which of the following algorithms **may result** in clusters like this?
Mark all applicable answers (1 - n completely filled boxes).

2/4

- [x] CHAMELEON
- [ ] PAM
- [ ] K-means
- [x] DBSCAN

Frage 31
Given are the **clustering features** CF_1, CF_2 and CF_3 for a small dataset:

$CF_1 = (2, (3, 1), (5, 1))$
$CF_2 = (2, (7, 3), (25, 5))$
$CF_3 = (2, (4, 7), (8, 25))$



Which values are stored in the **clustering feature** CF **combining** the above mentioned clustering features?
Mark one applicable answer (1 completely filled box).

0/4

- [ ] $(8, (84, 21), (1000, 125))$
- [ ] $(2, (3, 1), (5, 1))$
- [ ] $(2, (7, 7), (25, 25))$
- [ ] $(6, (84, 21), (1000, 125))$
- [x] $(6, (14, 11), (38, 31))$
- [ ] $(2, (4.67, 3.67), (12.67, 10.33))$

[redacted]

Outlier (Processing time approx. 6 min)

Frage 32
Which of the following statements are **True** or **False** with regard to the **types of outliers**.
Mark one applicable answer per statement (3 completely filled boxes in total).

0/3

● True ☒ False	Behavioral attributes are used to model contextual outliers such as to pick up noise in an attribute `temperature`.
■ True ☐ False	Local outliers are data points which density significantly deviates from the local area in which they occur.
● True ☒ False	Collective outliers are data points that deviate significantly to the rest of the dataset. Each individual data point in this group therefore is a global outlier.

Frage 33
Which of the following statements regarding **approaches to outlier detection** are *correct*?
Mark all applicable answers (1 - n completely filled boxes).

2/6

- ■ Grubb's test and Mahalanobis distance are examples of parametric statistical approaches to outlier detection.
- ● Non-parametric methods determine models from data without setting any parameters.
- ■ Maximum likelihod estimation is used to estimate parameters of parametric distribution such as gaussian distribution.
- ☐ Clustering methods like k-means models outliers as a separate cluster.
- ☐ A normal distribution can be used to find 68% of all outliers in a dataset.
- ● Using a mixture of parametric distributions for outlier detection, such as when using two normal distributions, detects outliers in a multivariate dataset, i. e. using one method per attribute.
- ■ Novelty detection finds outlier that exhibit new behaviour that differs to the already known outlier behaviours.

Frage 34
Which of the following statements are **True** or **False** with regard to **density-based outlier detection**.
Mark one applicable answer per statement (3 completely filled boxes in total).

0/3

☒ True ● False	Compared to density-based clustering, density-based outlier detection does not require to specify a minimum number of points in a neighborhood.
● True ☒ False	Density-based outlier detection methods detects global outliers.
☐ True ■ False	A point is considered an outlier when it is out of range of a pre-defined reachability distance, i. e. cannot be reached.
