Introduction (Processing time approx. 5.5 min)

Question 1

Data mining is defined as the extraction of interesting patterns from huge amounts of data.
With respect to this definition, which of the following characteristics **should** a pattern have to be considered interesting?
Mark all applicable answers (1 - n completely filled boxes).

1/2

☒ Potentially useful
□ Explicit

■ Non-specific
☒ Non-trivial

Question 2

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

0/4

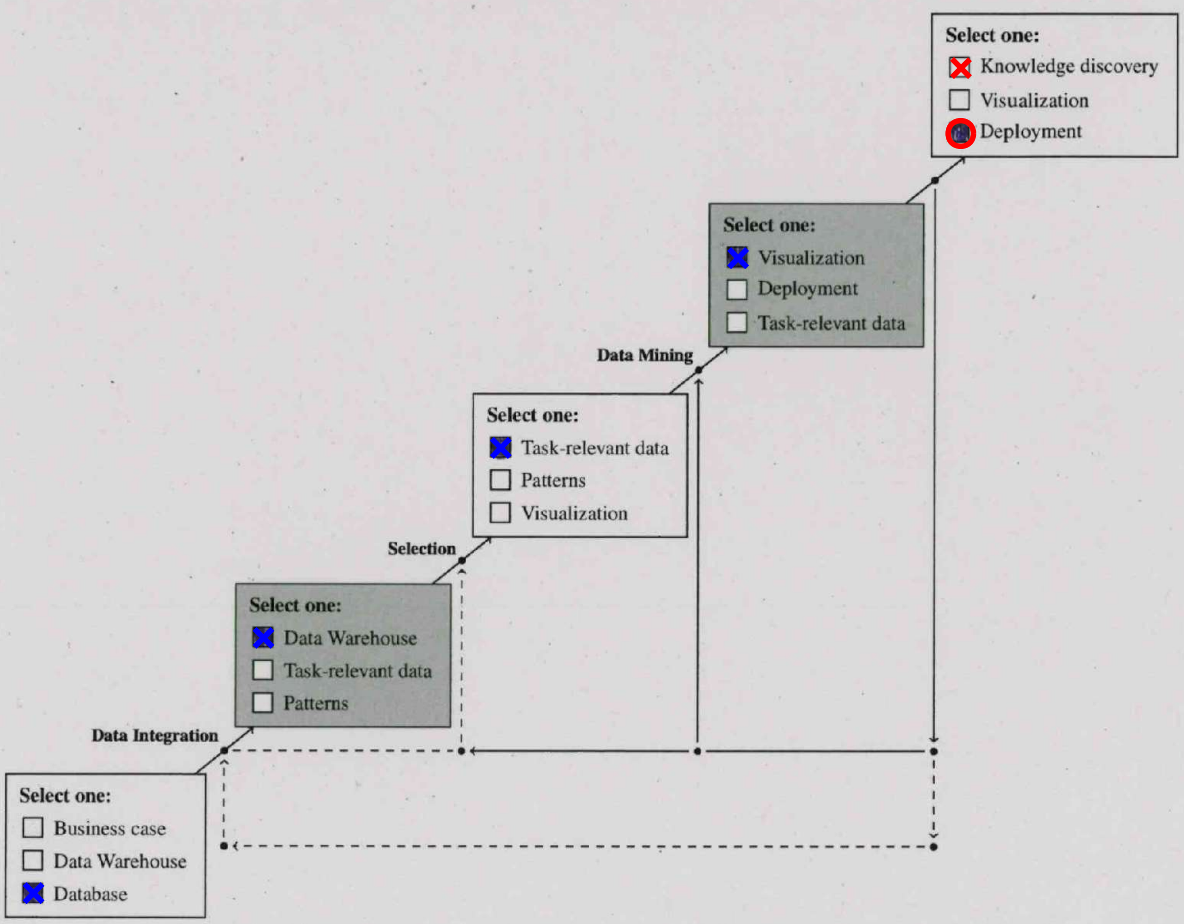
☒ **True** ■ **False** All kinds of data can be mined as long as they are meaningful for the target application.

☒ **True** □ **False** Machine learning, algorithms and statistics play a role in data mining.

Question 3

Given is the outline of the knowledge discovery pipeline.
Mark the terms that **are part of** this typical view of the database-systems community on data mining.
Mark one applicable answer per selection (5 completely filled boxes in total).

3/5



Select one:
□ Business case
□ Data Warehouse
☒ Database

Select one:
☒ Data Warehouse
□ Task-relevant data
□ Patterns

Select one:
☒ Task-relevant data
□ Patterns
□ Visualization

Select one:
☒ Visualization
□ Deployment
□ Task-relevant data

Select one:
☒ Knowledge discovery
□ Visualization
■ Deployment

Data (Processing time approx. 9 min)

Question 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 (4 completely filled boxes in total).

4/4

☒ **True** □ **False** Symmetric binary attributes are a specialization of discrete attributes.

□ **True** ☒ **False** Values of a nominal attribute have a meaningful order.

☒ **True** □ **False** Temperature in Kelvin and the masured length of an object are examples of ratio-scaled attributes.

☒ **True** □ **False** Quantitative measurements are often represented in integers or real values.

Question 5

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

8/10

- □ Sampling refers to the process of randomly splitting a dataset into training and test.
- ☒ It is possible that an attribute has no definite mode.
- ☒ A quantile-quantile plot compares the probability distributions of two attributes by plotting their respective quantiles against each other.
- ☒ The range is affected by extreme values.
- ● A dataset is a collection of all possible data objects and is therefore also called a population.
- ☒ A scatter plot can be used to visually detect if and how two attributes correlate with each other.
- ☒ The median of interval grouped data lies in the group which may first exceed 50% of the cumulative relative frequency.

Question 6

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

0/4

● **True** ☒ **False** A data matrix stores the distances of all possible object pairs.

☒ **True** □ **False** Similarity and dissimilarity are used in methods such as clustering, outlier analysis, and nearest-neighbor classification.

☒ **True** ● **False** Symmetry and the triangle inequality are among the desired properties of a dissimilarity function.

● **True** ✗ **False** Euclidean distance is a dissimilarity function for nominal and numerical data.

Preprocessing (Processing time approx. 16 min)

Question 7

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

2/8

– ☒ **True**	■ **False**	Small measurement inaccuracies are dirty data.
– □ **True**	☒ **False**	With incomplete input data, a data mining project always fails.
☒ **True**	□ **False**	Data scrubbing describes the use of domain knowledge to detect and correct errors.
☒ **True**	□ **False**	Binning with data smoothing is a measure to consider when dealing with noisy data.

Question 8

Given are **two** contingency tables, one for attributes **A and B** and one for attributes **C and D**:

	A_1	A_2	
B_1	20(10)	10(20)	30
B_2	15(25)	25(15)	40
	35	35	70

Attributes: A and B

	C_1	C_2	
D_1	90(100)	155(145)	245
D_2	210(200)	95(105)	305
	300	250	550

Attributes: C and D

$\Sigma \frac{c_{ij} - e_{ij}}{e_{ij}}$

$\frac{20-10}{10} + \frac{10-20}{20}$

$\frac{15-25}{25} + \frac{25-15}{15}$

Which of the attribute pairs is **more likely** to be related?
Mark one applicable answer (1 completely filled box).

6/6

☒ Attributes A and B
□ Attributes C and D
□ Both pairs are equally likely to be related
□ Both pairs are definitely not related

Question 9

What type of correlation is indicated if **Pearsons´s product-moment coefficient** has a value of **0.632**?
Mark one applicable answer (1 completely filled box).

3/3

□ Negative correlation
☒ Positive correlation
□ Correlation, but without tendency
□ Uncorrelated/no correlation

Question 10

Given is a **mystery normalization** function:

```
def mystery_normalization(df):
  return df / 10 ** (np.ceil(np.log10(df.abs().max())))
```

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

0/3

□ Min-Max normalization
■ Abs-Max normalization
☒ Normalization by decimal scaling
□ Z-score normalization

Question 11
Given is an **initial vector**:

$$V = (2, 3, 1, 9, 9, 5, 0, 6)$$

What vector represents the **detail coefficient** for a resolution of **4** if the **discrete wavelete transform** is applied on V?
Mark one applicable answer (1 completely filled box).

4/4

□ $(0.5, 4, -2, 3)$ □ $(1.25, -2)$ □ $(-1.25, 2)$ □ $(2.5, 5, 7, 3)$
□ $(3.75, 5)$ □ $(-2.5, -5, -7, -3)$ ☒ $(-0.5, -4, 2, -3)$ □ $(-3.75, -5)$

Question 12
Given are the **eigenvalues** and **eigenvectors** from a **principle components analysis** run-through:

$$\lambda_1 = +2.89, \nu_1 = \begin{bmatrix} +0.57 \\ +0.71 \\ +0.41 \\ -0.03 \end{bmatrix} \quad \lambda_2 = +0.00, \nu_2 = \begin{bmatrix} +0.57 \\ -0.71 \\ +0.41 \\ -0.03 \end{bmatrix} \quad \lambda_3 = +0.28, \nu_3 = \begin{bmatrix} +0.38 \\ +0.00 \\ -0.47 \\ +0.80 \end{bmatrix} \quad \lambda_4 = +0.83, \nu_4 = \begin{bmatrix} -0.45 \\ -0.00 \\ +0.66 \\ -0.60 \end{bmatrix}$$

Which of the **eigenvectors** should be selected to preserve at least **85%** of the original information? As **few** eigenvectors as possible should be chosen.
Mark one applicable answer per eigenvector (4 completely filled boxes in total).

4/4

☒ **Select** □ **Drop** ν_1 □ **Select** ☒ **Drop** ν_3
□ **Select** ☒ **Drop** ν_2 ☒ **Select** □ **Drop** ν_4

Question 13
Given is a **sorted list** of temperature values:

$$[21, 78, 81, 85]$$

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

2/4

☒ **Bin 1** □ **Bin 2** 21 □ **Bin 1** ☒ **Bin 2** 81
☒ **Bin 1** ■ **Bin 2** 78 □ **Bin 1** ☒ **Bin 2** 85

OLAP (Processing time approx. 5 min)

Question 14
Which of the following statements regarding **data darehouse in general** are **True**?
Mark all applicable answers (1 - n completely filled boxes).

☐ A data warehouse is optimized for massive write and read operations.
4/4 ☒ Data warehousing refers to the process of constructing and using a data warehouse.
☒ A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data.

Question 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 (4 completely filled boxes in total).

☐ **True** ☒ **False** In a star schema, a dimension exists for each concept hierarchy level.
☒ **True** ☐ **False** Fact tables contain measures and references to dimension tables.
4/4
☐ **True** ☒ **False** Data marts are sets of views over operational databases.
☒ **True** ☐ **False** A data cube allows a multi-dimensional view of data.

Question 16
For the following statements, mark whether they are **True** or **False** with regard to the **usage of a data warehouse in general**.
Mark one applicable answer per statement (2 completely filled boxes in total).

2/2 ☒ **True** ☐ **False** Elimination of duplicates is one task faced while integrating data.
☒ **True** ☐ **False** Operations like roll up and drill down utilize the underlying concept hierarchies of involved dimensions.

Frequent Patterns (Processing time approx. 15.5 min)

Question 17
Given are **all** itemsets for a dataset and their respective occurence counts:

Ski	5
Sticks	4
Helmet	2

Ski, Sticks	3
Ski, Helmet	2

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 (2 completely filled boxes in total)

4/4

☐ **True** ☒ **False** Given a minimum support count of 3, *Ski* would be a max-itemset.

☒ **True** ☐ **False** Given a minimum support count of 3, *Sticks* would be a closed itemset.

Question 18
Given is the **transactional dataset shown on the right**. Use **A Priori** to find all frequent itemsets for a minimum support count of 2.

Important: The frequent itemsets and **all** intermediate steps **have to** be written down.

TID	Items bought
1	*CPU, RAM*
2	*RAM*
3	*MB, FAN, CPU*
4	*MB, CPU, RAM*

Only for grading (do not fill in!): ☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ■9 ☐10

9/10

Intermediate steps:

min. sup : 2

①

Itemps	Counts
CPU	3
RAM	3
MB	2 ✓
FAN	1

drop FAN as |FAN| < min_sup →

CPU
RAM
MB ✓

→

Items	Count.
CPU, RAM	2
CPU, MB	2
RAM, MB ✓	1 ✓

drop RAM, MB →

CPU, RAM ✓
CPU, MB

→

Item	Count
CPU, RAM, MB	1

f1 Not a candidate as subset {RAM, MB} is not frequent

✓ Terminate

Frequent itemsets:

'CPU', 'RAM', 'MB' ✓, 'CPU, RAM' ✓, 'CPU, MB' ✓

Question 19
For the following statements, mark whether they are **True** or **False** with regard to **A Priori**.
Mark one applicable answer per statement (3 completely filled boxes in total).

□ **True** ☒ **False** The database is scanned a maximum of five times during a run.

6/6 ☒ **True** □ **False** Dynamic Itemset Counting may result in a smaller number of scans.

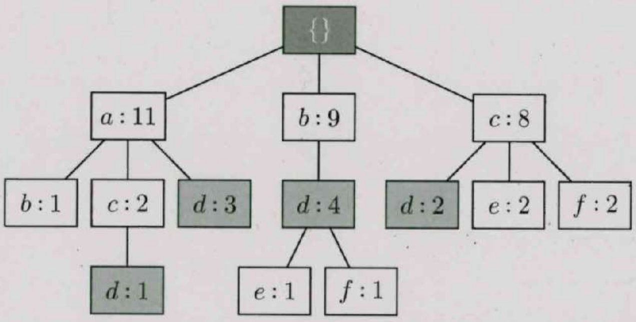
☒ **True** □ **False** Hashing is an efficient way to reduce the number of candidates.

Question 20
What is the **maximum** number of times the transactional database is scanned during a run of **FP-Growth**?
Mark one applicable answer (1 completely filled box).

3/3 □ 0 □ 1 ☒ 2 □ 3 □ 4 □ >=5

Question 21
Given is the **initial FP-Tree** from an **FP-Growth** run-through:

a:3 b:4
ac:1 c:2



Which of the following itemset/count combinations **belong to** the conditional pattern base of d?
Mark all applicable answers (1 - n completely filled boxes).

6/8

☒ $a : 3$
□ $cf : 8$
□ $ac : 2$

□ $e : 1$
□ $ab : 1$
□ $f : 1$

□ $b : 9$
☒ $b : 4$
☒ $c : 2$

□ $a : 11$
□ $c : 8$
☒ $ac : 1$

Classification (Processing time approx. 20 min)

Question 22
For the following statements, mark whether they are **True** or **False** with regard to **decision tree induction** in general.
Mark one applicable answer per statement (4 completely filled boxes in total).

□ **True** ☒ **False** Decision trees support continuous-valued attributes only when discretized beforehand.

■ **True** ☒ **False** An attribute selection method is a heuristic that ranks the list of attributes in a dataset in increasing order.

2/4

□ **True** ☒ **False** Decision tree induction is the process of learning a flowchart-like structure in a bottom-up manner.

□ **True** ☒ **False** A decision tree constructed with Gini index guarantees a balanced binary tree.

Question 23
A **decision tree with Gini index** should be constructed on a dataset.
For an attribute `B` with three possible values {`low`, `medium`, `high`}, **determine which statement is the correct split**.
Mark one applicable answer (1 completely filled box).

0/3

- □ Split on {`high`} and {`low`, `medium`}, with Gini index 0.0638.
- ☒ Split on {`low`} and {`medium`, `high`}, with Gini index 0.0630.
- ■ Split on {`low`}, {`medium`}, and {`high`}, with Gini index 0.0629.
- □ Split on {`medium`} and {`low`, `high`}, with Gini index 0.0635.

Question 24
You are about to buy a new laptop for university. Two laptop models sparked your particular interest, yet you cannot decide. Therefore, you consult your friends on what they would decide on. Naturally, each of your friend has different requirements for such a laptop. You gather each opinion and formulate your final decision based on a weighted combination of each friend's interest, requirement, recommended choice, as well as your own take on the laptop's suitability for your studies and personal use.

In which of the following **classification concepts** matches this procedure best?
Mark one applicable answer (1 completely filled box).

3/3 □ Random Forest □ Decision Tree □ Bagging ☒ Boosting

Question 25
For the following statements, mark whether they are **True** or **False** with regard to **model evaluation**.
Mark one applicable answer per statement (4 completely filled boxes in total).

☒ **True** □ **False** .632 bootstrap assigns a data tuple with an probability of 63.2 % to the training dataset.

□ **True** ☒ **False** For each class, stratified cross-validation samples data with replacement.

3/4

□ **True** ☒ **False** The F_1 measure combines precision and recall in one measure where precision is assigned more weight than recall.

□ **True** ☒ **False** ROC curve compares and shows the trade-off between the true positive rate and true negative rate.

Question 26

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:

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

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	medium	low	no
26	Raichu	low	high	no
78	Rapidash	medium	medium	no
136	Flareon	low	medium	no
144	Articuno	medium	medium	yes
145	Zapdos	medium	medium	yes
146	Moltres	medium	medium	yes
150	Mewtwo	medium	high	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.

9/13

Only for grading (do not fill in!):

□0 □1 □2 □3 □4 □5 □6 □7 □8 ■9
□10 □11 □12 □13

Calculation: ① Calculating Priors

$P(y = yes) = \frac{4}{10} = 0.4$ ✓ $P(y = No) = \frac{6}{10} = 0.6$ ✓

② Calculating likelihood

$P(\overset{f}{\cancel{X}}: defense = medium | y = yes) = \cancel{\frac{3}{6}} = \cancel{0.5}$ $\frac{4}{6}$ f

$P(\cancel{X}: defense = medium | y = No) = \cancel{\frac{3}{6}} = \cancel{0.25}$ $\frac{2}{6}$ f

$P(\cancel{X}: Speed = high | y = yes) = \frac{1}{2}$ f

$P(\cancel{X}: speed = high | y = No) = \frac{1}{2}$ f

0.133
$\frac{\cancel{0.04}}{\cancel{31}}$

③ Calculating Posterior

$P(y_{yes} | x) = P(\cancel{X}_{def} | y = yes) \times P(\cancel{X}_{speed} | y_{yes}) \times P(yes)$ V3 FF

$= \frac{4\,\cancel{2}\,1}{3\,\cancel{6}} \times \frac{1}{2} \times 0.4 = \boxed{0.1333}$

$P(y = No | x) = P(\cancel{X}_{def} | y = No) \times P(\cancel{X}_{spe} | y = No) \times P(No) = \frac{\cancel{2}}{6\,1} \times \frac{1}{\cancel{2}} \times \overset{0.1}{\cancel{0.6}} = 0.1$ V3FF

The Pokemon is most likely: Legendary with a probability of 13.38% VFF

Question 27

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

□ **True** ☒ **False**	Random forests use boosting and random attribute selection.	
☒ **True** □ **False**	A random forest classifier and AdaBoost achieve comparable accuracies.	4/4
☒ **True** □ **False**	Boosting assigns a weight to each training tuple.	
□ **True** ☒ **False**	A boosting classifier is usually faster constructed as a random forest classifier.	

The upcoming **three questions** refer to the following confusion matrix:

		Predicted Class		
		True	False	
True Class	True	30 TP	10 FP	40
	False	20 FP	40 TN	

$\frac{30+40}{100} = \frac{70}{100}$ 0.7

$\frac{TP}{P} = \frac{3\cancel{0}}{4\cancel{0}}$ 0.6

$\frac{TP}{TP+FP} = \frac{3\cancel{0}}{5\cancel{0}}$

Question 28

Calculate the metric **accuracy** and mark the appropriate value below.
Mark one applicable answer (1 completely filled box).

3/3

□ 0.30	□ 0.50	□ 0.60	□ 0.66	□ 0.75	□ 0.80
□ 0.33	□ 0.55	□ 0.65	☒ 0.70	□ 0.76	□ 0.86

Question 29

Calculate the metric **sensitivity** and mark the appropriate value below.
Mark one applicable answer (1 completely filled box).

3/3

□ 0.30	□ 0.50	□ 0.60	□ 0.66	☒ 0.75	□ 0.80
□ 0.33	□ 0.55	□ 0.65	□ 0.70	□ 0.76	□ 0.86

Question 30

Calculate the metric **precision** and mark the appropriate value below.
Mark one applicable answer (1 completely filled box).

3/3

□ 0.30	□ 0.50	☒ 0.60	□ 0.66	□ 0.75	□ 0.80
□ 0.33	□ 0.55	□ 0.65	□ 0.70	□ 0.76	□ 0.86

Clustering (Processing time approx. 14 min)

Question 31
Which of the following clustering methods **are among** the **partitioning** approaches?
Mark all applicable answers (1 - n completely filled boxes).

3/3 ☒ k-means ☐ CLIQUE ☐ DBSCAN ☒ PAM ☒ CLARA ☐ BIRCH

Question 32
Given are the coordinates of some **points**:

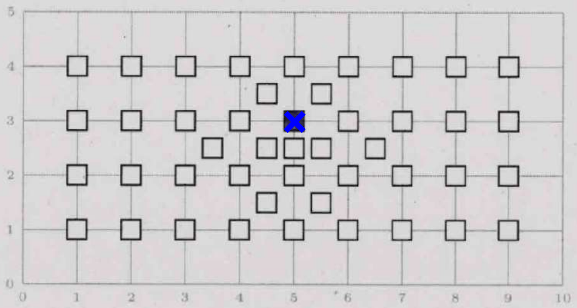
16 9 $\frac{25}{5} = 5$ $\frac{15}{5} = 3$

$$(1, 4), (2, 2), (5, 2), (8, 3), (9, 4)$$

What is the location of the **centroid** of these points? Mark the location **in** the coordinate system below.
Mark one applicable answer (1 completely filled box).

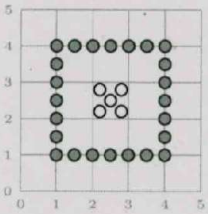
(5,3)

5/5

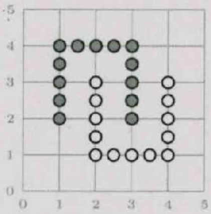


Question 33
Which of the following clustering results **might be** generated by **k-means**?
Mark all applicable answers (1 - n completely filled boxes).

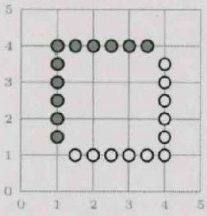
Legend: ● Cluster 1 ○ Cluster 2



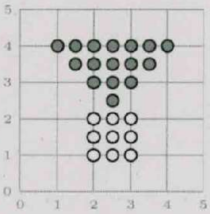
☐



☐



☒



☒

2/4

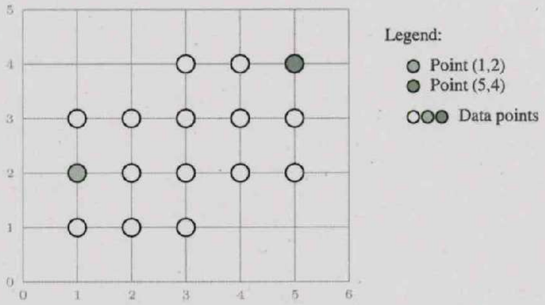
Question 34
For the following statements, mark whether they are **True** or **False** with regard to **hierarchical clustering**.
Mark one applicable answer per statement (2 completely filled boxes in total).

0/4

☐ **True** ☒ **False** A single-linkage algorithm is terminated based on the distance between the most distant clusters.

■ **True** ☒ **False** In DIANA (Divisive Analysis) the nodes that have the least dissimiliarity are merged.

Question 35
Given is a **coordinate system** with some **data points**:



For the following statements, assume that the **Euclidean distance** is used and mark whether they are **True** or **False** with regard to **density-based clustering**.
Mark one applicable answer per statement (3 completely filled boxes in total).

☒ **True** □ **False** Given $\epsilon = 1.1$ and $MinPts = 4$, $(1,2)$ and $(5,4)$ are density-connected.

2/6 ☒ **True** ● **False** Given $\epsilon = 1.1$ and $MinPts = 4$, $(5,4)$ is density-reachable from $(1,2)$.

□ **True** ☒ **False** Given $\epsilon = 1.1$ and $MinPts = 4$, $(1,2)$ is density-reachable from $(5,4)$.

Question 36
For the following statements, mark whether they are **True** or **False** with regard to the **evaluation of clustering**.
Mark one applicable answer per statement (3 completely filled boxes in total).

☒ **True** □ **False** Splitting a small cluster into pieces is more harmful than splitting a large cluster.

2/6 ☒ **True** □ **False** The Hopkins statistic is a way of measuring the cluster tendency of a dataset.

● **True** ☒ **False** Extrinsic methods can be used when the ground truth is not available.

Outlier (Processing time approx. 5 min)

Question 37
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).

□ **True** ☒ **False** Contextual outliers can be viewed as a specialization of local outliers.

2/3 ☒ **True** □ **False** A large set of stock transactions of the same stock among a small group of brokers in a short period of time can be viewed as a collective outlier.

□ **True** ☒ **False** Global outliers can be viewed as contextual outliers where the set of contextual attributes is empty.

Question 38
Which of the following statements are **True** or **False** with regard to **approaches to outlier detection**.
Mark one applicable answer per statement (7 completely filled boxes in total).

□ **True** ☒ **False** Histogram is a simple parametric method that can be used to detect outliers.

□ **True** ☒ **False** Both k-medoid and k-means model outliers in one dedicated dense cluster.

☒ **True** ■ **False** Clustering-based approaches require small manageable computational costs to model and then find outliers.

1/7 ■ **True** ☒ **False** Noise generally deviates significantly from normal data objects and could be detected as outliers with statistical methods.

■ **True** ☒ **False** Statistical approaches to outlier detection assume that outliers are generated by some unknown distribution.

□ **True** ☒ **False** Kernel Density Estimation can be used to estimate appropriate bin sizes of a histogram.

☒ **True** □ **False** Grubb's test is a two-sided test to detect outliers in an univariate data set.
