Online Test 2 (10%)

- Due 12 Nov at 16:00
- Points 100
- Questions 21
- Available 12 Nov at 12:45 - 13 Nov at 3:59 15 hours and 14 minutes
- Time limit 60 Minutes

Instructions

Due Date: Week 11 Lab Session

Weighting: 10%

About this test

1. **You must participate in this test in person on campus.**
2. This is an online test worth 10% of the total marks for this unit.
3. It consists of 20 questions on theories and applications. Using KNIME in the quiz is essential to find answers.
4. You may **only attempt this test once**.
5. **Do not navigate away from the test (close the browser tag) before you have completed it and submitted it.**
6. You have **60 minutes** to complete the test; exceeding this will result in a loss of marks and penalties.
7. This is a **close-book test**.
8. Once you begin the quiz, a summary will appear in the upper right corner, showing the number of questions (completed and total) and the time remaining.
9. **Please bring your student ID to the test and put it on the desk in front of you so the tutor can verify your identity.**
10. **Please use Chrome to access the quiz. The second options are Firefox and Microsoft Edge. Please do not use Safari in the test.**
11. **The multi-answering questions have a deduction mechanism. Be careful about what you select in the answers.**

Attempt history

	Attempt	Time	Score
LATEST	Attempt 1	60 minutes	82.5 out of 100

Score for this quiz: 82.5 out of 100

Submitted 12 Nov at 14:01

This attempt took 60 minutes.

Question 1

0 / 5 pts

Which of the following is NOT a part of the problem framing process in the data discovery phase?

Correct answer

- Identify the key stakeholders.

You Answered

- Establish failure criteria.
- Identify the main objectives of the project.
- Write down the problem statement and share it with the key stakeholders.

Week 07 - Lecture - P20.

Question 2

5 / 5 pts

Which one on the list is NOT one of the considered factors when deciding the data store?

Correct!

- Collect the data from data sources.
- Identify the goals.
- Avoid Data Fatigue.
- Data Volume. (Big or Small Data)

Week 07 - Lecture - P102.

Question 3

5 / 5 pts

Which of the following methods is (are) the discretisation technique(s)?

- Association Rule
- Classification

Correct!

- Histogram

Correct!

- Binning

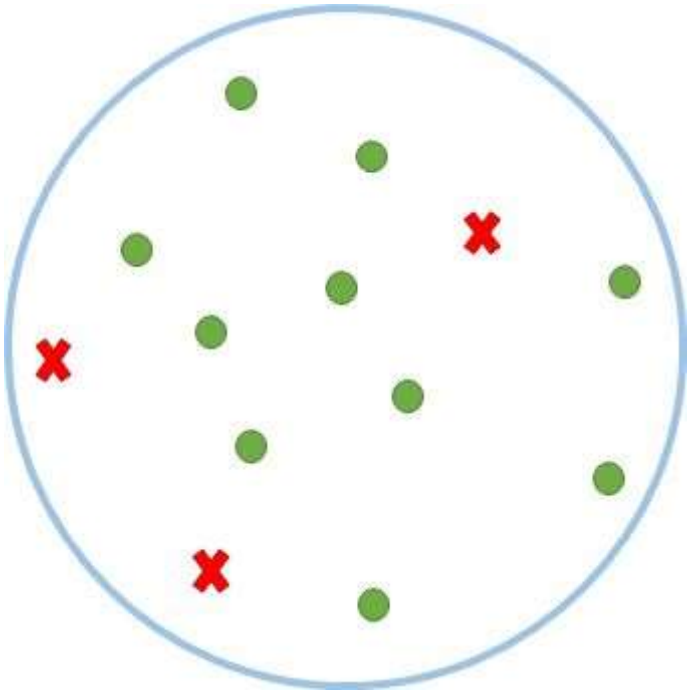
Week 08 - Lecture - P54.

Question 4

5 / 5 pts

Consider the given figure below.

Which of the following is (are) correct?



Note that $\log_2(x) = \frac{\log_{10}(x)}{\log_{10}(2)}$

Correct!

- [x] The entropy of the dataset is 0.78

Correct!

- [x] The Gini impurity of the dataset is 0.36
- [ ] The entropy of the dataset is 0.48
- [ ] The Gini impurity of the dataset is 0.88

$p_{circle} = \frac{10}{13}$

$p_{cr} = \frac{3}{13}$

$H(p) = -\frac{10}{13}\log_2(\frac{10}{13}) - \frac{3}{13}\log_2(\frac{3}{13}) \approx 0.78$

$Gini_{impurit} = 1 - \frac{3^2}{13^2} - \frac{10^2}{13^2} \approx 0.36$

Question 5

5 / 5 pts

If you find a numerical column in the dataset, whose mean, median, and mode values are identical, the data distribution in this column must be ...

- ( ) Multi-peak distribution.
- ( ) Negatively skewed distribution.

Correct!

- (●) Normal distribution.
- ( ) Positively skewed distribution.

Week 08 - Lecture - P11.

Question 6

5 / 5 pts

If you want to assign known labels to objects, what will possibly be the best tool to use in most cases?

Correct!

- (●) Classification
- ( ) Association Rules
- ( ) Time-series Analysis
- ( ) Clustering Methods

Week 07 - Lecture - P4.

Question 7

5 / 5 pts

Which of the following is (are) correct?

Correct!

- [x] The Random forest model utilises multiple trees to reduce the risk of overfitting.
- [ ] Random forest runs based on odd number of trees only.

Correct!

- [x] Random forest can maintain accuracy when a large proportion of data is missing.
- [ ] Random forest runs inefficiently on large dataset.

Week 05 - Lecture - P41.

Question 8

5 / 5 pts

Which one on the list belongs to the 3-dimensional data?

- ( ) Text

Correct!

- (●) Trajectory
- ( ) Audio
- ( ) Photo

Week 07 - Lecture - P98.

Question 9

5 / 5 pts

Which of the following methods is (are) commonly used in ensemble learning?

Correct!

- [x] Stacking
- [ ] Wrangling

Correct!

- [x] Bootstrapping

Correct!

- [x] Boosting

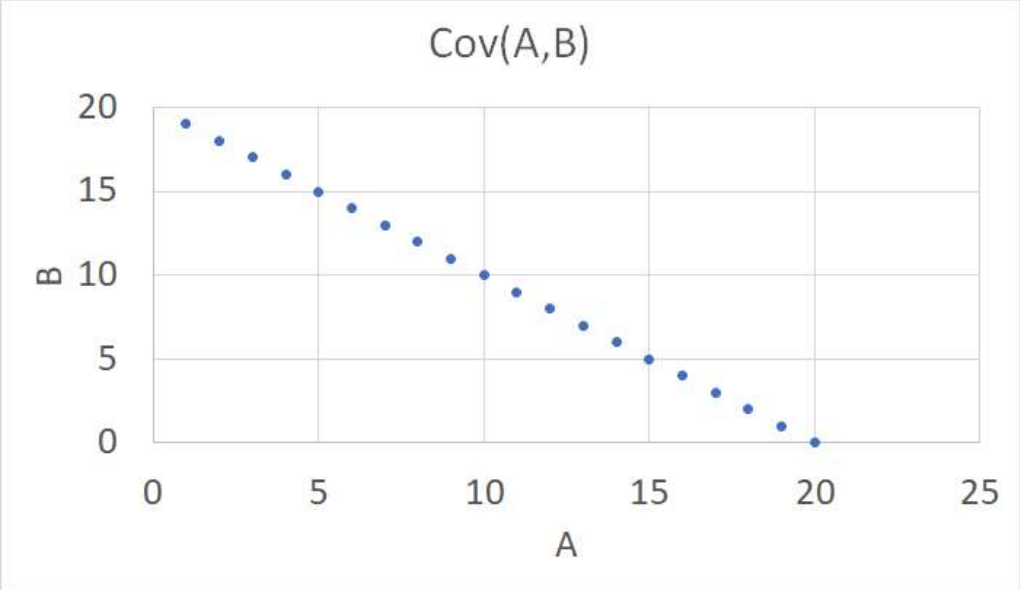
Week 05 - Lecture - P25.

Question 10

5 / 5 pts

The figure below shows the covariance (Cov) for two numeric attributes A and B.

Based on this figure, which one of the description is true?



- [ ] We can't tell the relationship of covariance from the given figure.
- [ ] Attributes A and B have near zero covariance.

Correct!

- [x] Attributes A and B have a negative covariance.
- [ ] Attributes A and B have a positive covariance.

Week 08 - Lecture - P30.

Question 11

5 / 5 pts

Which of the following has the correct explanation to HiPPO Effect?

- It means that the person speaking the loudest in the data science team would drive the direction of analysis.

Correct!

- It means that the authority figure's suggestions are interpreted as the final truth, and promptly implemented, even if the findings from the data are contrary.
- It means to build the strongest foundation for your theory before doing any analysis.
- It means to pack your theory as strong as a hippo can make your claim beats the others.

Week 07 - Lecture - P16.

Question 12

5 / 5 pts

What are the key thresholds used in the Apriori algorithm for discovering Association Rules?

- support and assumption.
- confidence and liverage.

Correct!

- support and confidence.
- confidence and lift.

Week 06 - Lecture - P11.

Question 13

5 / 5 pts

Which of the following descriptions regarding Apriori Property (downward closure property) is correct?

Correct!

- If an item is considered frequent, then any subset of the frequent itemset must also be frequent.
- If an item is considered infrequent, then any subset of the frequent itemset is possibly to be frequent.
- If an item is considered frequent, then any subset of the frequent itemset must not be frequent.
- If an item is considered infrequent, then any subset of the frequent itemset must be frequent.

Week 06 - Lecture - P16.

Question 14

5 / 5 pts

In the Data Analytics Lifecycle, which of the following is the correct sequence?

Correct!

- Discovery -> Data Preparation-> Model Planning -> Model Building -> Communicate Results -> Operationalise.
- Data Preparation -> Discovery -> Model Planning -> Model Building -> Operationalise -> Communicate Results.
- Data Preparation -> Discovery -> Model Planning -> Model Building -> Communicate Results -> Operationalise.
- Discovery -> Model Planning -> Data Preparation-> Model Building -> Communicate Results -> Operationalise.

Week 07 - Lecture - P8.

Question 15

0 / 5 pts

Which of the following descriptions is (are) NOT the proper way(s) to handle missing data?

- [ ] Use a global constant to fill in the missing values.
- [ ] Fill in the missing value with side information.

Correct answer

- [ ] Randomly pull a value from the same column in other tuples to fill in the missing value.

You Answered

- [x] Ignore the tuple.

Week 08 - Lecture - P10.

Question 16

5 / 5 pts

Given the table below. Which criterion (criteria) will be the most appropriate for splitting the dataset in classification based on the brand?

S/N	Dimension	Colour	Weight	Brand
1	20	Red	154	A
2	30	Green	147	B
3	32	Brown	139	B
4	26	Pink	162	A
5	24	Yellow	188	A
6	28	Black	114	B
7	29	Orange	138	B

- [ ] Colour?

Correct!

- [x] Dimension > 27?
- [ ] S/N is odd?

Correct!

- [x] Weight < 150?

Week 05 - Lecture - P16.

Given the purchase transaction dataset below, try to process the data with KNIME or any preferred tool to answer the questions.

Dataset: **Transactions.csv (https://swinburne.instructure.com/courses/62961/files/33246093?wrap=1)** **(https://swinburne.instructure.com/courses/62961/files/33246093/download?download_frd=1)**

Question 17
5 / 5 pts

We plan to apply data mining to the given dataset to find out possible hidden relationships between goods.

Assume our support and confidence values are 20% and 70%, respectively.

The maximum number of rules to be found is set to 300.

Try to build a workflow in KNIME or use any available tool to find the answers.

Which of the following description(s) is (are) TRUE?

- [ ] Rule 4 says that a customer buys the egg is highly possible to buy the wine, the chicken, and the cheese along with the egg.
- [ ] The rule of a customer buying the cheese and the apple will also buy the chicken is more important the the rule of a customer buys the chicken, the cheese, and the eggs will also buy the wine.

Correct!

- [x] The lift value of the second rule is higher than the first rule in the mining result.

Correct!

- [x] 50 rules in total are found based on the given support and confidence criteria.

Question 18
2.5 / 5 pts

We plan to apply data mining to the given dataset to find out possible hidden relationships between goods.

Assume our support and confidence values are 15% and 60%, respectively.

The maximum number of rules to be found is set to 300.

Try to build a workflow in KNIME or use any available tool to find the answers.

Which of the following description(s) is (are) TRUE?

- [ ] If a customer buys the wine, the cheese, and the apple, there is above 85% chance that the customer will also buy the egg.

Correct answer

- [ ] If a customer buys the wine, the chicken, the egg, and the apple, there is above 85% chance that the customer will also buy the cheese.

Correct!

- [x] The mining result returns with one L5 large itemset.
- [ ] 300 association rules are found with the given support and confidence criteria.

Question 19

2.5 / 5 pts

We plan to apply data mining to the given dataset to find out possible hidden relationships between goods.

Assume our support and confidence values are 40% and 20%, respectively.

Try to build a workflow in KNIME or use any available tool to find the answers.

Which of the following description(s) is (are) TRUE?

Correct!

- [x] The large itemset only contains a pair of vegatable and apple.
- [ ] Two L2 item pairs are found in the result.

Correct!

- [x] Only two available rules are obtained by the given support and confidence criteria.

You Answered

- [x] The probability of someone buys the apple also buys the vegatable is greater or equal to someone buys the vegatble also buys the apple.

Question 20

2.5 / 5 pts

We plan to apply data mining to the given dataset to find out possible hidden relationships between goods.

Assume our support and lift values are 20% and 90%, respectively.

The maximum number of rules to be found is set to 300.

Try to build a workflow in KNIME or use any available tool to find the answers.

Which of the following description(s) is (are) TRUE?

Correct!

- [x] There are 18 rules that their lift values are above or equal to 1.5
- [ ] Both rule 1 and 2 have the same lift value and thus their confidence values are also the same.
- [ ] Rule 1 and 2 are equavelent because all item combinations in them are identical.

Correct answer

- [ ] There is no L5 itemsets available under the given criteria.

Quiz score: 82.5 out of 100