Question 1

Given the items $I = \{A, B, C, D, E, F, G, H, I\}$
and the set of transactions T:

TransID	Items
1	A B C E G H I
2	A B D E F H I
3	A B D E H
4	A B E F H
5	A B E H
6	A D F G I
7	A F I
8	B C D E G I
9	C G I
10	D E F G H I
11	D G I
12	F

For the minimum support of 3, we already determined the frequent 3-itemsets with the APRIORI algorithm:

$$L_3 = \{ABE, ABH, AEH, AFI, BDE, BEH, BEI, CGI,$$
$$DEH, DEI, DFI, DGI, EFH, EGI, EHI\}$$

Which of the following 4-itemsets are preliminary candidates in the next step of APRIORI (i.e., after the merging step but before pruning)?

1. ABDE
2. ABEH
3. BEHI
4. CDGI
5. CEGI
6. DEFI
7. DEGI
8. DEHI

9. EFHI
10. EGHI

Solution:
Generate candidate set C_4 using L_3. Condition of joining two itemsets is that they should have $(k-2)$ elements in common, which is 2 in this case.

Candidate set C_4 would be

$$L_4 = \{ABEH, BEHI, DEHI\}$$

Thus itemsets 2, 3, 8 are preliminary candidates in step 4.

Question 2

For some transaction database we found that the rule $\{A, B, C, D\} \Rightarrow \{E, F, G\}$ has a confidence below the confidence threshold.

Which of the following rules will therefore have a confidence below the confidence threshold as well?

1. $\{A\} \Rightarrow \{B, C, D, E, F, G\}$
2. $\{A, B, C, D\} \Rightarrow \{E, F\}$
3. $\{A, C\} \Rightarrow \{B, E, F\}$
4. $\{A, C\} \Rightarrow \{B, D, E, F, G\}$
5. $\{A, D\} \Rightarrow \{B, E, G\}$
6. $\{A, D\} \Rightarrow \{B, E, F, G\}$
7. $\{A, D\} \Rightarrow \{B, C, E, F, G\}$
8. $\{B, E, F\} \Rightarrow \{A, C, D\}$
9. $\{C\} \Rightarrow \{A, B, D, E, F, G\}$
10. $\{C, D\} \Rightarrow \{A, B, E, F, G\}$

Solution:
Recall when we have an association rule $x \Rightarrow y$.
Support is the amount of times x and y appear together.
Confidence is the support, divided by the times that the first part of the expression appears alone in the table, i.e. $\frac{s}{|x|}$.

1. $\underline{\{A\} \Rightarrow \{B, C, D, E, F, G\}}$
   Support would be the same.
   The amount of times A would appear in the table would be more than, or equal to, the amount of times $\{A, B, C, D\}$ appear in the table.
   Thus, the s in confidence would be the same, but the amount that it would be divided by would be equal or more. So in conclusion it would be below the confidence threshold.

2. $\{A, B, C, D\} \Rightarrow \{E, F\}$
   Support would be more than or equal, since the total amount of items is smaller.
   The amount of times $\{A, B, C, D\}$ is in the table is the same.
   Thus, in order to calculate the confidence, s would be more than or equal, and the bottom would be the same. So it would not be guaranteed to be below the confidence threshold.

3. $\{A, C\} \Rightarrow \{B, E, F\}$
   Support would be more than or equal, since the total amount of items is smaller.
   The amount of times $\{A, C\}$ would appear in the table would be more than, or equal to, the amount of times $\{A, B, C, D\}$ appear in the table.
   Thus, in order to calculate the confidence, s would be more than or equal, and the bottom also be more. So it would not be guaranteed to be below the confidence threshold.

4. $\underline{\{A, C\} \Rightarrow \{B, D, E, F, G\}}$
   Support would be the same.
   The amount of times $\{A, C\}$ would appear in the table would be more than, or equal to, the amount of times $\{A, B, C, D\}$ appear in the table.
   Thus, the s in confidence would be the same, but the amount that it would be divided by would be equal or more. So in conclusion it would be below the confidence threshold.

5. $\{A, D\} \Rightarrow \{B, E, G\}$
   It would not be guaranteed to be below the confidence threshold.

6. $\{A, D\} \Rightarrow \{B, E, F, G\}$
   It would not be guaranteed to be below the confidence threshold.

7. $\underline{\{A, D\} \Rightarrow \{B, C, E, F, G\}}$
   Support would be the same.
   The amount of times $\{A, D\}$ would appear in the table would be more than, or equal to, the amount of times $\{A, B, C, D\}$ appear in the table.
   Thus, the s in confidence would be the same, but the amount that it would be divided by would be equal or more. So it would be below the confidence threshold.

8. $\{B, E, F\} \Rightarrow \{A, C, D\}$
   It would not be guaranteed to be below the confidence threshold.

9. $\underline{\{C\} \Rightarrow \{A, B, D, E, F, G\}}$
   Support would be the same.
   The amount of times $\{C\}$ would appear in the table would be more than, or equal to, the amount of times $\{A, B, C, D\}$ appear in the table.
   Thus, the s in confidence would be the same, but the amount that it would be divided by would be equal or more. So it would be below the confidence threshold.

10. $\underline{\{C, D\} \Rightarrow \{A, B, E, F, G\}}$
    Support would be the same.
    The amount of times $\{C, D\}$ would appear in the table would be more than, or equal to, the amount of times $\{A, B, C, D\}$ appear in the table.
    Thus, the s in confidence would be the same, but the amount that it would be divided by would be equal or more. So it would be below the confidence threshold.

Question 3

We have the following one-dimensional dataset:

ID	Value
A	1
B	3
C	5
D	7
E	10
F	11
G	12

In three attempts, k-means delivered the following three clustering solutions:

$$S_1 = \{A, B, C\}, \{D, E, F, G\}$$
$$S_2 = \{A, B\}, \{C, D\}, \{E, F, G\}$$
$$S_3 = \{A, B, C, D\}, \{E, F, G\}$$

We want to compare the solutions using TD^2. Which of the following statements are correct?

1. S_1 is better than S_2 in terms of TD^2.
2. S_2 is better than S_3 in terms of TD^2.
3. S_1 and S_3 are equally good in terms of TD^2.
4. S_3 is better than S_1 in terms of TD^2.

Solution:
TD^2 is a measure of compactness for a cluster. Recall how TD^2 is calculated for a cluster C

$$TD^2(C) = \sum_{p \in C} \text{dist}(p, \mu_C)^2$$

We want to calculate the centroids for the three clusterings, and then their respective TD^2 values.

$\underline{S_1}$:

$$\begin{aligned}
\mu_1 &= \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 3 + \frac{1}{3} \cdot 5 \\
&= 3 \\
\mu_2 &= \frac{1}{4} \cdot 7 + \frac{1}{4} \cdot 10 + \frac{1}{4} \cdot 11 + \frac{1}{4} \cdot 12 \\
&= 10 \\
TD^2(1) &= 2^2 + 0^2 + 2^2 = 8 \\
TD^2(2) &= 3^2 + 0^2 + 1^2 + 2^2 = 14 \\
TD^2 &= 8 + 14 = 22
\end{aligned}$$

$\underline{S_2}$:

$$\begin{aligned}
\mu_1 &= \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 3 \\
&= 2 \\
\mu_2 &= \frac{1}{2} \cdot 5 + \frac{1}{2} \cdot 7 \\
&= 6 \\
\mu_3 &= \frac{1}{3} \cdot 10 + \frac{1}{3} \cdot 11 + \frac{1}{3} \cdot 12 \\
&= 11 \\
TD^2(1) &= 1^2 + 1^2 = 2 \\
TD^2(2) &= 1^2 + 1^2 = 2 \\
TD^2(3) &= 1^2 + 0^2 + 1^2 = 2 \\
TD^2 &= 2 + 2 + 2 = 6
\end{aligned}$$

$\underline{S_3}$:

$$\begin{aligned}
\mu_1 &= \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 3 + \frac{1}{4} \cdot 5 + \frac{1}{4} \cdot 7 \\
&= 4 \\
\mu_2 &= \frac{1}{3} \cdot 10 + \frac{1}{3} \cdot 11 + \frac{1}{3} \cdot 12 \\
&= 11 \\
TD^2(1) &= 3^2 + 1^2 + 1^2 + 3^2 = 20 \\
TD^2(2) &= 1^2 + 1^2 = 2 \\
TD^2 &= 20 + 2 = 22
\end{aligned}$$

The correct statements are statement 2 and 3.

Question 4

ID	forecast	humidity	wind	play tennis?
1	sunny	high	weak	no
2	sunny	high	strong	no
3	sunny	high	weak	yes
4	sunny	normal	weak	yes
5	sunny	normal	strong	no
6	rainy	high	weak	no
7	rainy	normal	weak	yes
8	rainy	normal	weak	yes
9	rainy	normal	strong	yes
10	rainy	high	strong	no

A decision tree is being trained on the above data set. As root of the tree, the attribute "forecast" was already selected.

Which attributes are selected as test nodes at the next level based on the Gini index?

1. For the branch of forecast=sunny, we test wind.
2. For the branch of forecast=sunny, we test humidity.
3. For the branch of forecast=rainy, we test wind.
4. For the branch of forecast=rainy, we test humidity.

Solution:
We choose the attribute and the split that minimizes the Gini index.

For forecast=sunny:
$|T| = 5$, 3 "no" and 2 "yes".

- G(humidity)

- high: T_1 = persons 1, 2, 3

$$\begin{aligned} p(PT = no) &= \frac{2}{3} \\ p(PT = yes) &= \frac{1}{3} \\ G(T_1) &= 1 - \left(\frac{2^2}{3^2} + \frac{1^2}{3^2}\right) \\ &= \frac{4}{9} \end{aligned}$$

- normal: T_2 = persons 4, 5

$$\begin{aligned} p(PT = no) &= \frac{1}{2} \\ p(PT = yes) &= \frac{1}{2} \\ G(T_2) &= 1 - \left(\frac{1^2}{2^2} + \frac{1^2}{2^2}\right) \\ &= \frac{1}{2} \end{aligned}$$

Thus, we can calculate the Gini index

$$G(\text{humidity}) = \frac{3}{5} \cdot \frac{4}{9} + \frac{2}{5} \cdot \frac{1}{2} = \frac{7}{15} = 0.47$$

- G(wind)
  - weak: T_1 = persons 1, 3, 4

$$\begin{aligned} p(PT = no) &= \frac{1}{3} \\ p(PT = yes) &= \frac{2}{3} \\ G(T_1) &= 1 - \left(\frac{1^2}{3^2} + \frac{2^2}{3^2}\right) \\ &= \frac{4}{9} \end{aligned}$$

- strong: T_2 = persons 2, 5

$$\begin{aligned} p(PT = no) &= \frac{2}{2} \\ p(PT = yes) &= \frac{0}{2} \\ G(T_2) &= 1 - \left(\frac{2^2}{2^2} + \frac{0^2}{2^2}\right) \\ &= 0 \end{aligned}$$

Thus, we can calculate the Gini index

$$G(\text{wind}) = \frac{3}{5} \cdot \frac{4}{9} + \frac{2}{5} \cdot 0 = \frac{4}{15} = 0.27$$

- Since $G(\text{wind}) < G(\text{humidity})$, we choose to split on wind for forecast=sunny.

For forecast=rainy:
$|T| = 5$, 2 "no" and 3 "yes".

- G(humidity)
  - high: T_1 = persons 6, 10

$$\begin{aligned} p(PT = no) &= \frac{2}{2} \\ p(PT = yes) &= \frac{0}{2} \\ G(T_1) &= 1 - \left(\frac{2^2}{2^2} + \frac{0^2}{2^2}\right) \\ &= 0 \end{aligned}$$

  - normal: T_2 = persons 7, 8, 9

$$\begin{aligned} p(PT = no) &= \frac{0}{3} \\ p(PT = yes) &= \frac{3}{3} \\ G(T_2) &= 1 - \left(\frac{0^2}{3^2} + \frac{3^2}{3^2}\right) \\ &= 0 \end{aligned}$$

Thus, we can calculate the Gini index

$$G(\text{humidity}) = \frac{2}{5} \cdot 0 + \frac{3}{5} \cdot 0 = 0$$

- $G(\text{wind})$
  - weak: $T_1 =$ persons 6, 7, 8

$$\begin{aligned} p(PT = no) &= \frac{1}{3} \\ p(PT = yes) &= \frac{2}{3} \\ G(T_1) &= 1 - \left(\frac{1^2}{3^2} + \frac{2^2}{3^2}\right) \\ &= \frac{4}{9} \end{aligned}$$

  - strong: $T_2 =$ persons 9, 10

$$\begin{aligned} p(PT = no) &= \frac{1}{2} \\ p(PT = yes) &= \frac{1}{2} \\ G(T_2) &= 1 - \left(\frac{1^2}{2^2} + \frac{1^2}{2^2}\right) \\ &= \frac{1}{2} \end{aligned}$$

  Thus, we can calculate the Gini index

$$G(\text{wind}) = \frac{3}{5} \cdot \frac{4}{9} + \frac{2}{5} \cdot \frac{1}{2} = \frac{4}{15} = \frac{7}{15} = 0.47$$

- Since $G(\text{humidity}) < G(\text{wind})$, we choose to split on humidity for forecast=rainy.

Thus we can conclude that the correct statements are 1 and 4.

Question 5

In a dataset with ten points $\{A, B, C, D, E, F, G, H, I, J\}$, A and B are labeled outliers.

Four outlier detection methods, $m_1, \ldots, m_4$, deliver the following rankings (from left-to-right: top-rank to bottom-rank):

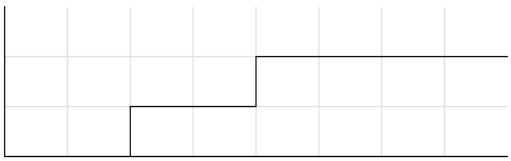
method	ranking
m_1	C,D,A,E,F,B,G,H,I,J
m_2	J,A,D,E,F,G,B,H,I,C
m_3	I,D,A,E,F,G,B,H,C,J
m_4	I,J,E,A,B,F,G,H,C,D

Based on ROC AUC as evaluation measure, which of the following statements is correct?

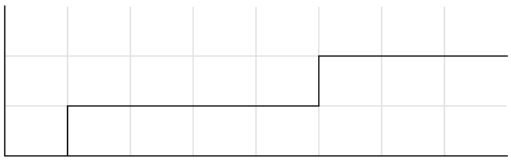
1. m_1 and m_2 perform equally well.
2. m_2 is better than m_3.
3. m_3 is better than m_4
4. m_2 is better than m_4

Solution:
We can create the ROC AUC for each detection method.

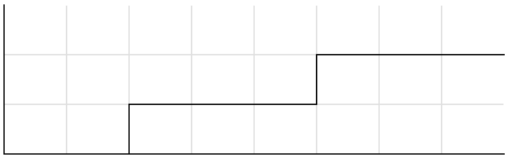
For m_1: $\frac{10}{24} \approx 0.41667$



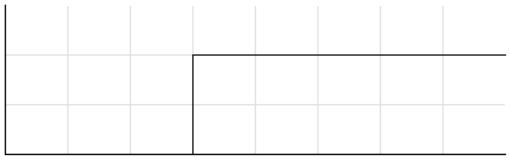
For m_2: $\frac{10}{24} \approx 0.41667$



For m_3: $\frac{9}{24} = 0.375$



For m_4: $\frac{10}{24} \approx 0.41667$



Thus we can see that the correct statements are statements 1 and 2.