ECS766P-210764484-Elliot Linsey

Q1

A

1. Predicting sales trends using past data can be considered a data mining task. This is because we are looking for interesting patterns that can be used to influence decisions made by the store. We could potentially use methods such as autoregression (AR) or an autoregressive moving average model (ARMA).
2. This is not a data mining task as this is more a data entry task. We are not analysing the data, simply inputting the values and calculating the sum. For it to be data mining we would have to be trying to find patterns or useful information from the data which we are not doing in this instance.

B

- Patient ID: Qualitative Categorical, nominal.
- Age: Quantitative Numeric, ratio scaled
- Admitted to Emergencies: Qualitative Categorical, binary asymmetric (being admitted to emergency is more important than not)
- Patient Risk: Qualitative Categorical, ordinal

C

This is a sparse data matrix, the data objects can be thought of as points (vectors) in a multidimensional space, where each dimension represents a distinct attribute describing the object. As this is a sparse data matrix the attributes are asymmetric and only non-zero values are important.

D

Manhattan

absolute(2-3) = 1

absolute(2-5) = 3

Manhattan distance = 3+1 = 4

Supremum

absolute(2-3) = 1

absolute(2-5) = 3

max distance = 3

Supremum = 3

E

If the scale is:

- extremely dissatisfied: 1
- somewhat dissatisfied: 2
- neutral: 3
- somewhat satisfied: 4
- extremely satisfied: 5

The dissimilarity (D) between ES and SD is 5-2/(n-1) = 0.75

Similarity = 1 - D

Similarity = 0.25

Q2

A

As several attributes are correlated, we could use Principal Component Analysis (PCA) to convert these into linearly uncorrelated variables called principal components. In doing this, we project the data onto a much smaller space, resulting in dimensionality reduction.

B

1. For studying enrolments for the past decade, we could use a data warehouse system utilising OLAP (Online Analytical Processing). This is because the data is not being used in day to day operations and can be stored independently.
2. For registering students to modules it could use DBMS (Database Management System with OLTP (Online transactional processing). This is used for day to day situations and is based more on transactions, which in this case would be assigning modules to students.

C

We should use z-score normalisation. This is because there is an outlier of 36 within the data that will dominate if we used min-max normalisation.

D

We can calculate the correlation coefficient:

Mean of age = 47

Mean of SBP = 128.83

Standard deviation of age = 17.27

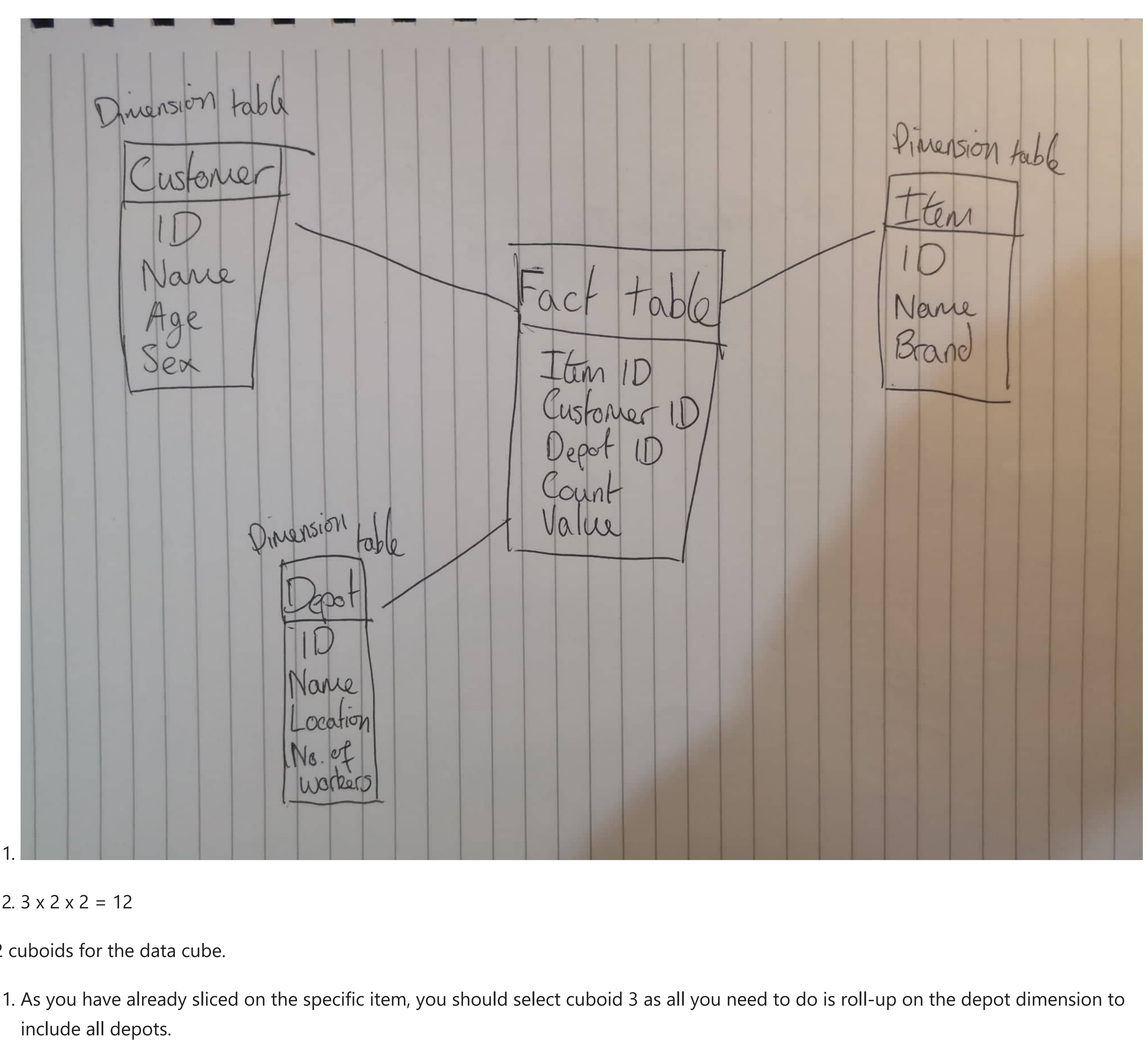
Standard deviation of SBP = 9.00

$\Sigma(a_i b_i)$ = 37144

$r = \frac{(37144 - (6 \cdot 47 \cdot 128.83)}{6 \cdot 17.27 \cdot 9.00}$
$r = 0.87$

As r = 0.87, we can confirm that both Age and SBP are positively correlated

E

1. 

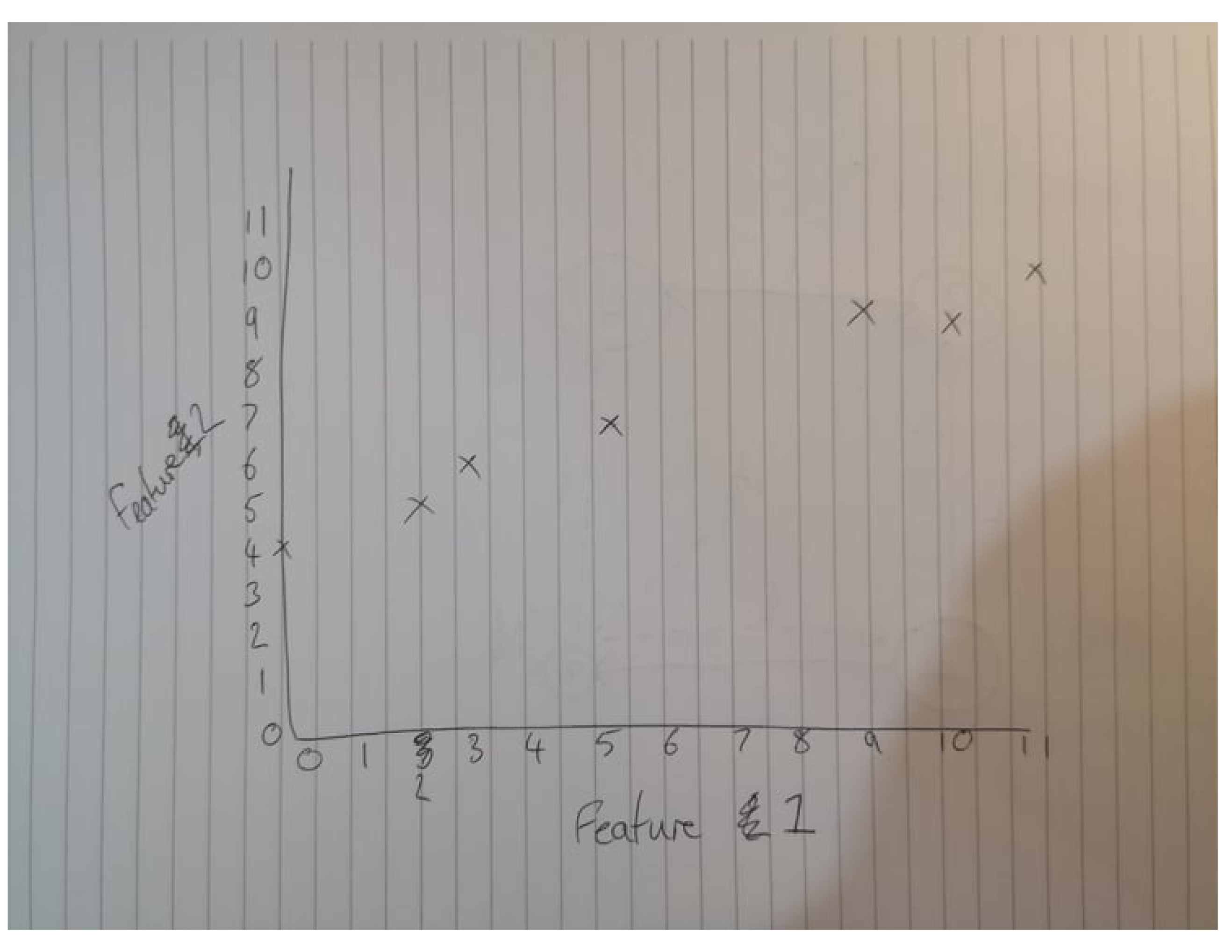
2. 3 x 2 x 2 = 12

12 cuboids for the data cube.

1. As you have already sliced on the specific item, you should select cuboid 3 as all you need to do is roll-up on the depot dimension to include all depots.

Q3

A



1. You can state the correlation coefficient will be close to 1 as the features appear positively correlated.

2.

- 0: 2
- 1: 2
- 2: 2
- 3: 2

The dataset has no mode as all the numbers appear the same amount of times.

B

1. If a feature has a large range in a distance-based classifier compared to other features, it will dominate the results as they are not on the same scale as the other features. Comparing age to income would skew the classifier towards making income the most important as that ranges in the thousands whereas age only ranges from 0 to 100.
2. No, the test set should be the final step in the evaluation process. You should not change any parameters or dimensions after using the test set as this defeats its purpose. Parameter and feature selection should have been explored using the training and validation sets.

C

1:

Distance from (0.5,0.5) to x1 (0.5,1):

((0.5 - 0.5)^2 + (0.5 - 1)^2)^0.5 = 0.5

Distance from (0.5,0.5) to x2 (1,0.5):

((0.5 - 1)^2 + (0.5 - 0.5)^2)^0.5 = 0.5

Distance from (0.5,0.5) to x3 (1,1):

((0.5 - 0.5)^2 + (0.5 - 1)^2)^0.5 = 0.71

The point would be classed as 1 as it is closest to both points x1 and x2 which both have the class 1.

2:

K-nearest neighbours needs a tie-breaking policy when the K value is even. This is because if a new point has 2 neighbours that are closest to it that both have different classes in a K=2 classification, then there is no majority class and therefore it has to be able to choose which class to select using a tie-breaking policy.

D

1: (0.5 + 1 + 1)/3 = 0.83

(1 + 0.5 + 1)/3 = 0.83

Cluster center = (0.83,0.83)

2: Minimising the SSE means that datapoints are closer to their respective cluster centroids. The more clusters we have, the lower the SSE and therefore the more dense our clustering. However, we need to find a trade-off between SSE and the number of clusters as if we have K = N number of clusters, the SSE would be 0 as each datapoint would be its own cluster and therefore have a distance of 0. Using the elbow method, we find the greatest drop in SSE and find where it starts to stabilise and select this value of K.

E

If a classifier has perfect accuracy then it also has perfect recall and precision

The formula for F1 score is $2 \cdot \frac{P \cdot R}{P + R}$

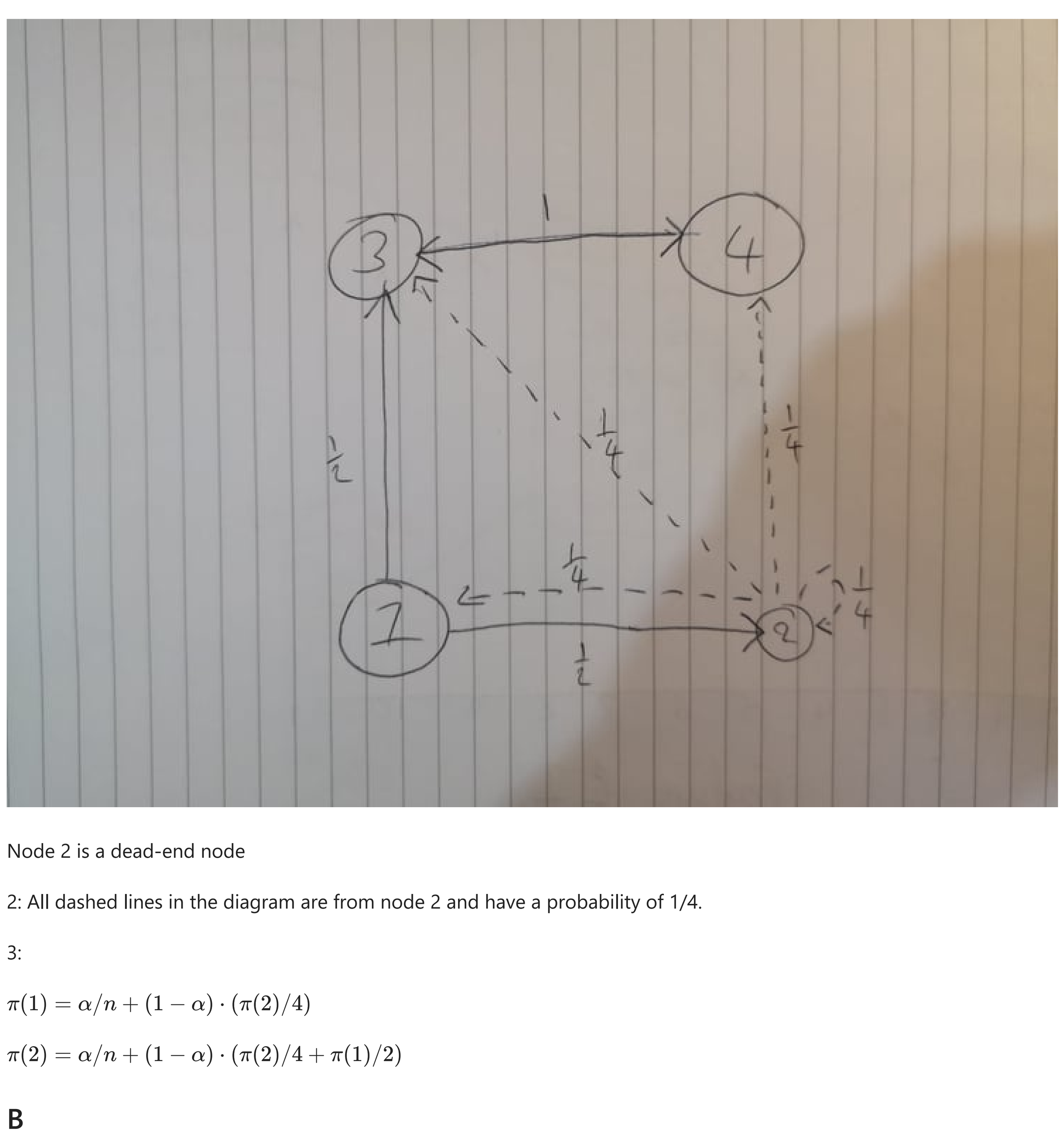
Therefore the perfect scores for precision and recall in this classifier =

$2 \cdot \frac{1 \cdot 1}{1+1}$
$= 1$

Q4

A

1:



Node 2 is a dead-end node

2: All dashed lines in the diagram are from node 2 and have a probability of 1/4.

3:

$\pi(1) = \alpha/n + (1 - \alpha) \cdot (\pi(2)/4)$

$\pi(2) = \alpha/n + (1 - \alpha) \cdot (\pi(2)/4 + \pi(1)/2)$

B

We perform frequent itemset mining to reduce the search space within our transaction database and itemset. The reason we look for strong association rules is to find items that are frequently purchased together. In order for us to find this we have to find all itemsets that are frequently purchased. Using the apriori itemset mining method means that we eliminate itemsets that are not frequently purchased and therefore we eliminate itemsets that have low association rules.

C

1:

Support count of {1,3} = 4

Support of {2,4} = 2/8 = 1/4

2:

Support of {1,3,4} = 1/4 = 25%

Therefore it is considered frequent for a threshold of 20%.

As {1,4} is a subset of {1,3,4}, according to the apriori method the support will be greater than or equal to the support of {1,3,4}, therefore {1,4} should also be frequent for a threshold of 20%.

3:

Support of $[1, 3] \rightarrow [4]$ = 1/4

Confidence = 2/4 = 1/2

D

1:

We could use a K means clustering algorithm to group data into clusters. We would then find the distance between each point and its cluster mean and assign it a distance based outlier score. By finding the average distance of all points to their cluster centre, we could then divide the specific point distance by the average distance. If this result is large, then the point is likely an outlier as it is greater than the average distance for all points in its respective cluster.

2:

Similarities:

- Both can use unlabeled data as they are both subsets of unsupervised learning.
- They can work for many types of data, including higher dimensions
- Both may be computationally expensive, it can be difficult to cluster large datasets.

Differences:

- Different ability to classify collective outliers, for K means a small group of outliers should still be far away from their centroid so can be classified easier. For density, this small group may be missed as the group density may give it a low outlier score.
- Density can capture local outliers due to the different densities in comparison to the local neighbourhoods, distance cannot capture these local outliers.