Chin Jin Yao, C130129@e.ntu.edu.sg

Question 1:

(a):

i. F
ii. F
iii. F
iv. F, chameleon is clustering using dynamic modeling, and clustering is a descriptive data mining task.
v. F, classification is a predictive data mining task.
vi. F
vii. F, all combinations of values are likely to occur, hence the data cube would be dense.
viii. F
ix. T
x. F, when all the individual classifiers produce identical results, and ensemble of such classifiers will be no different from using one of these classifiers, therefore the error rates will not be reduced.

(b):

(i):

Staff ID:	Discrete, Nominal
Gender:	Binary, Nominal
Age:	Discrete, Ratio
Pay:	Discrete, Ratio
Designation:	Discrete, Ordinal
Postcode:	Discrete, Nominal

(ii):

Knowing that gender is Male, designation is Engineer, and Age is 25, there are 3 other staff who are also male engineers.

Table 1: Male Engineers ordered by Age

Staff ID	Gender	Designation	Age	Pay
0003	Male	Engineer	34	\$3,000
0006	Male	Engineer	45	\$3,500
0005	Male	Engineer	55	\$4,000

The 3 male engineers have an age difference of roughly 10 years, and their pay difference is \$500. Based on that, given that the newly employed male engineer has an age of 25, his pay can be estimated to be $2,500 by using the available information.

There can be other ways to estimate his pay, so the solution is not unique.

(iii):

Mode: Engineer, Frequency: 3.

(iv): **Min-max normalization for <u>Pay</u> values**

Min: $2,500, Max: $9,000. [2500,9000] normalized to [0.0,1.0].

Staff ID 0005: $4,000
Normalized Pay: ($4,000 - $2,500) / ($9,000 - $2,500) = 0.231

Staff ID 0006: $3,500
Normalized Pay: ($3,500 - $2,500) / ($9,000 - $2,500) = 0.231

(v): **Min-max normalization for <u>Age</u> values**

Mean (μ): (45+58+34+27+55+45+46+23) / 8 = 41.625

Standard Deviation = 11.77

Staff ID 0005: 55
Normalized Pay: (55 – 41.625) / 11.77 = 1.136

Staff ID 0006: 45
Normalized Pay: (45 – 41.625) / 11.77 = 0.286

Question 2:

(a):

Overall Gini: $= 1 - \left(\frac{3}{8}\right)^2 - \left(\frac{5}{8}\right)^2 = 0.46875$

Customer ID: Customer ID is unique. For every customer ID, GINI index will be $1 - \left(\frac{1}{1}\right)^2$ which is 0. Gini$_{split}$ using Customer ID will be $\frac{1}{8} \times 0 + \frac{1}{8} \times 0 + \cdots + \frac{1}{8} \times 0 = 0$

Drive Car:

$$Gini(Drive\ Car = Yes) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

$$Gini(Drive\ Car = No) = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

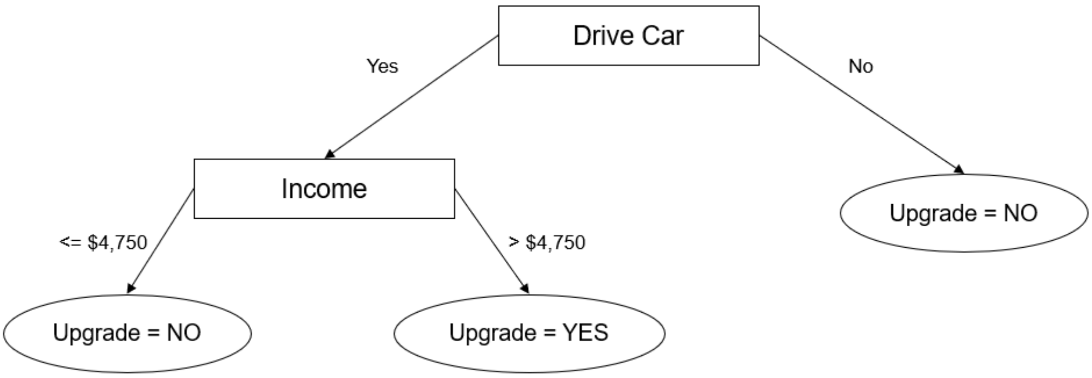
$$Gini(split) = \frac{5}{8} \times 0.48 + \frac{3}{8} \times 0 = 0.3$$

(ii):

Split	$2,250		$2,750		$3,250		$3,750		$4,250		$4,750		$5,250	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Upgrade = YES	0	3	0	3	1	2	1	2	2	1	2	1	3	0
Upgrade = NO	0	5	1	4	2	3	3	2	3	2	5	0	5	0
Gini	0.46875		0.4285		0.4666		0.4375		0.4666		0.3571		0.46875	

Best $Gini_{split}$ can be obtained using v = $4,750 where the GINI value is 0.3571.

(iii):



Based on 2(a)(i) and 2(a)(ii), at the first level, Drive Car is the best attribute with the most reduction in GINI.

At second level, repeat binarization for Income attribute by considering only customers with attribute Drive Car = Yes. The best split is still $4,750.

(iv):

$$Generalization\ Error\ of\ original\ data: \frac{3}{8} = 0.375$$

$$Generalization\ Error\ of\ tree: \frac{(2 + 3 \times 0.5)}{8} = 0.4375$$

The decision tree is not a better classifier.

(v):

If Drive Car = No, Upgrade = NO.
If Drive Car = Yes and Income <= $4,750, Upgrade = NO.

If Drive Car = Yes and Income > $4,750, Upgrade = YES.

For customer with Drive Car = No and Income = $5,000, Upgrade = NO.

(vi):

P(Upgrade = YES | Income = $5,000 and Drive Car = No)

$$= \frac{P(Income = \$5{,}000\ and\ Drive\ Car = No|Upgrade = YES)P(Upgrade = YES)}{P(Income = \$5{,}000\ and\ Drive\ Car = No)}$$

$$= \frac{P(Income = \$5{,}000|Upgrade = YES)(Drive\ Car = No|Upgrade = YES)P(Upgrade = YES)}{P(Income = \$5{,}000\ and\ Drive\ Car = No)}$$

Since $(Drive\ Car = No|Upgrade = YES) = 0$, P(Upgrade = YES | Income = $5,000 and Drive Car = No) will be 0. Therefore, for the customer with Income = $5,000 and Drive Car = No, the predicted class label of Upgrade will be NO.

Question 3:

(a):

(i):

Simple Matching Coefficient (SMC): (00+11) / (00+01+10+11)

Similarity Matrix (Symmetrical):

	D1	D2	D3	D4	D5
D1	0	0.6	0.5	0.5	0.5
D2		0	0.7	0.5	0.1
D3			0	0.6	0.4
D4				0	0.6
D5					0

Convert to Distance (or Dissimilarity) Matrix:

	D1	D2	D3	D4	D5
D1	0	0.4	0.5	0.5	0.5
D2		0	0.3	0.5	0.9
D3			0	0.4	0.6
D4				0	0.4
D5					0

(ii):

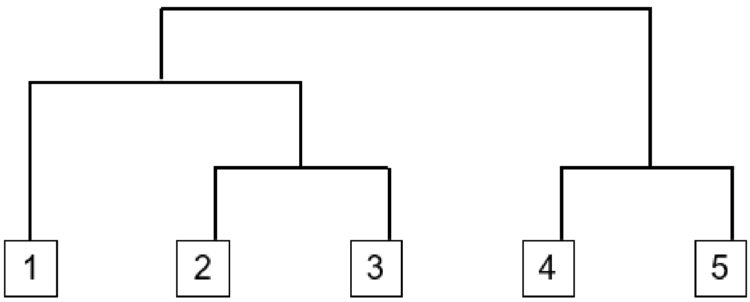
	D1	D2,3	D4	D5
D1	0	0.45	0.5	0.5
D2,3		0	0.45	0.75
D4			0	0.4
D5				0

For average-link, $distance(Ci, Cj) = \frac{\sum Pi \in Ci, Pj \in Cj \; distance(Pi, Pj)}{|Ci| \times |Cj|}$

Distance(D1, D2,3) = distance(D, D) + distance(D1, D2) / 1 * 2 = (0.4 + 0.5) / 2 = 0.45
Distance(D4, D2,3) = distance(D, D) + distance(D4, D2) / 1 * 2 = (0.5 + 0.4) / 2 = 0.45
Distance(D5, D2,3) = distance(D, D) + distance(D5, D2) / 1 * 2 = (0.9 + 0.6) / 2 = 0.75

Combine D4 with D5, which has distance of 0.4.
Recalculate distance matrix, Combine D1 with D2,3.
Recalculate distance matrix, Combine D1,2,3 with D4,5

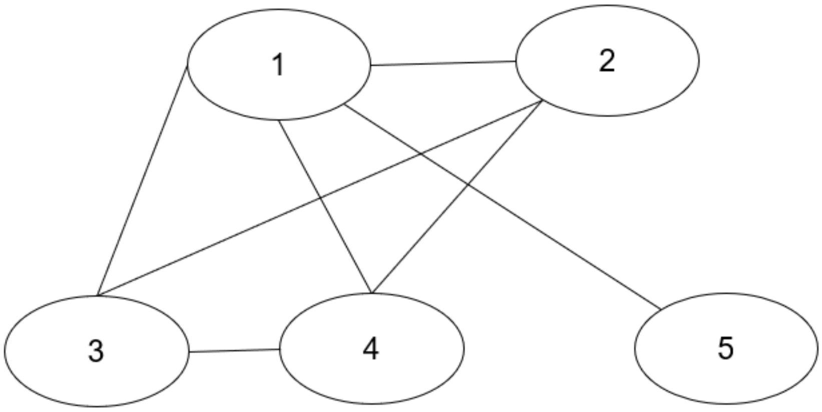
Dendrogram:



(iii):

	1NN	2NN	3NN
D1	D2	D3, D4, D5	-
D2	D3	D1	D4
D3	D2	D4	D1
D4	D3, D5	D1, D2	-
D5	D4	D1	D3

	D1	D2	D3	D4	D5
D1	0	2	2	3	2
D2		0	2	2	0
D3			0	2	0
D4				0	0
D5					0



Number of clusters: 1

(b):

DBSCAN involves finding the core points, border points, and noise points. A point is a core point if it has more than a specified number of points (MinPts) within Eps (including itself). A border point is not in a core point, but is in the neighborhood of a core point. A noise point is any point that is not a core point or a border point. The algorithm works by eliminating noise points and then performing clustering on the remaining points by placing two core points within Eps into the same cluster. After which border points are added to the nearest cluster.

(c):

CURE uses a number of points to represent a cluster. Representative points are found by selecting a constant number of points from a cluster, and then shrinking them towards the center of the cluster. Shrinking representative points toward the center helps to avoid problems with noise and outliers. It is better able to handle clusters of arbitrary shapes and sizes.

Question 4:

(a):

(i): Treating each <u>customer ID</u> as the <u>basic unit</u>, there are <u>3 transactions</u>.

Customer ID	Items Purchased
C001	{a, b, c, e}
C002	{a, b, c, d}
C003	{a, b, d, e}

Min_sup = 50%, and min_sup_count = 3*0.5 = 1.5.

1-candidates	
Itemset	Support
A	3
B	3
C	2
D	2
E	2

Freq 1-itemsets	
Itemset	Support
A	3
B	3
C	2
D	2
E	2

2-candidates	
Itemset	Support
AB	3
AC	2
AD	2
AE	2
BC	2
BD	2
BE	2
CD	1
CE	1
DE	1

Freq 2-itemsets	
Itemset	Support
AB	3
AC	2
AD	2
AE	2
BC	2
BD	2
BE	2

3-candidates	
Itemset	Support
ABC	2
ABD	2
ABE	2
ACD	1
ACE	1
BCD	1
BCE	1
BDE	1

Freq 3-itemsets	
Itemset	Support
ABC	2
ABD	2
ABE	2

4-candidates	
Itemset	Support
ABCD	1
ABCE	1
ABDE	1

No more frequent itemset

Min_conf = 80% or 0.8

Confidence of Association Rules:

C(A → B) = 3/3 , C(B → A) = 3/3, C(A → C) = 2/3, C(C → A) = 2/2, C(A → D) = 2/3, C(D → A) = 2/2, C(A → E) = 2/3, C(E → A) = 2/2, C(B → C) = 2/3, C(C → B) = 2/2, C(B → D) = 2/3, C(D → B) = 2/2, C(B → E) = 2/3, C(E → B) = 2/2

C(AB → C) = 2/3, C(AC → B) = 2/2, C(BC → A) = 2/2, C(A → BC) = 2/3, C(B → AC) = 2/3, C(C → AB) = 2/2, C(AB → D) = 2/3, C(AD → B) = 2/2, C(BD → A) = 2/2, C(A → BD) = 2/3, C(B → AD) = 2/3, C(D → AB) = 2/2, C(AB → E) = 2/3, C(AE → B) = 2/2, C(BE → A) = 2/2, C(A → BE) = 2/3, C(B → AE) = 2/3, C(E → AB) = 2/2

(ii):

Treating <u>each Transaction ID</u> as the <u>basic unit</u>, there are <u>7 transactions</u>.

min_sup = 50%, and min_sup_count = **7 *** 0.5 = 3.5

1-itemset	
Itemset	Support
A	5
B	5
C	5
D	2
E	4

→

Freq 1-itemsets	
Itemset	Support
A	5
B	5
C	5
E	4

→

2-itemset	
Itemset	Support
AB	4
AC	3
AE	3
BC	4
BE	3
CE	2

→

Freq 2-itemsets	
Itemset	Support
AB	4
BC	4

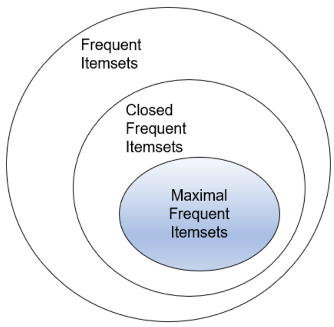
→

3-itemset	
Itemset	Support
ABC	3

→

No more frequent itemset

- Infrequent (Denoted by I)
- Closed (Denoted by C): *An itemset is closed if none of its immediate supersets has the same support as the itemset.*
- Maximal Frequent (Denoted by M): *An itemset is maximal frequent if none of its immediate supersets are frequent.*



<table>
<tr><td colspan="10">Null</td></tr>
<tr><td colspan="2">A: C</td><td colspan="2">B: C</td><td colspan="2">C: C</td><td colspan="2">D: I</td><td colspan="2">E: C, M</td></tr>
<tr><td>AB:C,M</td><td>AC: I</td><td>AD: I</td><td>AE: I</td><td>BC:C,M</td><td>BD: I</td><td>BE: I</td><td>CD: I</td><td>CE: I</td><td>DE: I</td></tr>
<tr><td>ABC: I</td><td>ABD: I</td><td>ABE: I</td><td>ACD: I</td><td>ACE: I</td><td>ADE: I</td><td>BCD: I</td><td>BCE: I</td><td>BDE: I</td><td>CDE: I</td></tr>
<tr><td colspan="2">ABCD: I</td><td colspan="2">ABCE: I</td><td colspan="2">ABDE: I</td><td colspan="2">ACDE: I</td><td colspan="2">BCDE: I</td></tr>
<tr><td colspan="10">ABCDE: I</td></tr>
</table>

Start by finding <u>all infrequent itemsets</u>, all infrequent itemsets <u>**cannot**</u> be closed/maximal. Find all closed itemsets, e.g. AB is <u>closed</u> because it has support count of 4, but ABC only has a support count of 3, and any other superset of AB will not have a support count of 4 either. Find all maximal frequent itemsets. AB is maximal frequent as all its immediate supersets ABC, ABD, and ABE are infrequent. AC is maximal frequent as all its immediate supersets ABC, ACD, and ACE are infrequent. E is maximal frequent as all its immediate supersets AE, BE, CE, and DE are infrequent.

Min_conf = 80% or 0.8

Confidence of Association Rule:

c(A →B) = 4/5, c(B →A) = 4/5, c(B →C) = 4/5, c(C →B) = 4/5,

(b):

Outliers are data points that are considerably different from remainder of damage.

The Grubbs' test is a statistical approach used to detect outliers in univariate data. Assuming that the data comes from a normal distribution, the Grubbs' test detects one outlier at a time, removes the outlier, and repeat the process.

H_0: There is no outlier in the data

H_A: There is at least one outlier

Grubbs' test statistic:

$$G = \frac{max|X - \bar{X}|}{s}$$

H_0 will be rejected, and the process will be repeated after removing the outlier, if

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2{}_{(\alpha/N, N-2)}}{N - 2 + t^2{}_{(\alpha/N, N-2)}}}$$

Amendments to answer key:

1(b)(v):

Standard Deviation = $\sqrt{158.267857143} = 12.5805$ using sample formula

Staff ID 0005: 55
Normalized Pay: (55 – 41.625) / 12.5805 = 1.06315

Staff ID 0006: 45
Normalized Pay: (45 – 41.625) / 12.5805 = 0.26827