Solver: Chua Wen Kai

1)

a)

i) Ans: **C**

For (i), it is just simple accounting.
From (ii), it is just signal processing.
For (iii), able to build a model of the normal behavior of heart rate and raise an alarm when an unusual heart behavior occurred. This would involve the area of data mining known as anomaly detection.

ii) Ans: **G**

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male*, *female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good*, *better*, *best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

This categorization of attributes is due to S. S. Stevens

Aside from nominal attribute type, the others all possess the properties order.

iii) Ans: **D**

Correlation between 2 data objects that have binary or continuous variables is a measure of the linear relationship between the attributes of the objects which can be calculated using Pearson's correlation coefficient which ranges from -1 to 1.

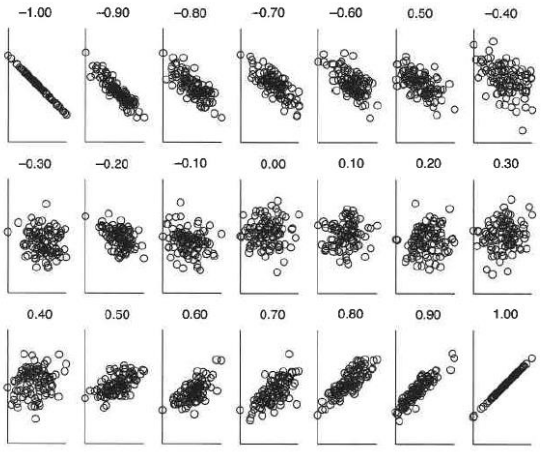


Figure 2.17. Scatter plots illustrating correlations from −1 to 1.

Hence, a correlation of 1(-1) means that x and y have a perfect positive (negative) linear

relationship.

iv) Ans: **E**
To perform feature reduction in the following steps which involves elimination of features with respect to their correlation to other features or use of regression to select features with respect to their corresponding weights.

v) Ans: **A**
Weak learners are sure about particular part of a problem. Hence, they usually don't overfit which means that weak learners have low variance and high bias.

vi) Ans: **B**

- Multiple runs
  - Helps, but probability is not on your side
- Sample and use hierarchical clustering to determine initial centroids
- Select more than k initial centroids and then select among these initial centroids
  - Select most widely separated
- Postprocessing
- Bisecting K-means
  - Not as susceptible to initialization issues

vii) Ans: **D**
Training error relates to the number of misclassification errors committed on training records, whereas validation error is the expected error of the model on previously unseen records. Good models will have low TE as well as low VE.

b) Both simple matching and jaccard coefficient are not applicable in this case as the vectors are not binary.
Euclidean distance (L_2 norm):

$$\left(\sum_{k=1}^{n}|p_k - q_k|^2\right)^{\frac{1}{2}} = \sqrt{(1-0)^2 + (0-1)^2 + \cdots + (1-0)^2 + (1-1)^2} = \sqrt{11}$$

Supremum distance (L_{max} norm):

$$\max(|p_1 - q_1|, |p_2 - q_2|, \ldots, |p_{n-1} - q_{n-1}||p_n - q_n|) = 2$$

Cosine Similarity:

$$\frac{(p \cdot q)}{\|p\|\|q\|} = \frac{11}{\sqrt{13}\sqrt{20}} \approx 0.6822$$

Correlation:

$$\mu_p = 0.7, \mu_q = 1, \sigma_p = 0.9487, \sigma_q = 1.0541$$

$$covariance(p,q) = \frac{1}{n-1}\left(\sum_{k=1}^{n}(p_k - \mu_p)(q_k - \mu_q)\right) = \frac{1}{9}(4) = \frac{4}{9}$$

$$\frac{covariance(p,q)}{std(p) \times std(q)} = \frac{4}{9 \times 0.9487 \times 1.0541} \approx 0.4444$$

c) Both outliers and anomalies are data points considerably different from the remainder of data, however, anomalies can be classified into three types of categories: point anomalies; contextual anomalies and collective anomalies.
There are no differences in their corresponding detection algorithms.

2)

a) Sorted values: 30,40,60,80,100

$$Gini(\leq 35) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0\,, Gini(> 35) = 1 - \left(\frac{7}{8}\right)^2 - \left(\frac{1}{8}\right)^2 = \frac{7}{32}\,,$$

$$\boldsymbol{Gini(split) = \frac{2}{10}(0) + \frac{8}{10}\left(\frac{7}{32}\right) = 0.175}$$

$$Gini(\leq 50) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48\,, Gini(> 50) = 1 - \left(\frac{5}{5}\right)^2 - \left(\frac{0}{5}\right)^2 = 0\,,$$

$$Gini(split) = \frac{5}{10}(0.48) + \frac{5}{10}(0) = 0.24$$

$$Gini(\leq 70) = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = \frac{24}{49}\,, Gini(> 70) = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0,$$

$$Gini(split) = \frac{7}{10}\left(\frac{24}{49}\right) + \frac{3}{10}(0) = 0.343$$

$$Gini(\leq 90) = 1 - \left(\frac{6}{9}\right)^2 - \left(\frac{3}{9}\right)^2 = \frac{4}{9}\,, Gini(> 90) = 1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2 = 0\,,$$

$$Gini(split) = \frac{9}{10}\left(\frac{4}{9}\right) + \frac{1}{10}(0) = 0.4$$

b)

$$Gini(parent) = 1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2 = 0.42$$

$$Gini(Own_{Condo} = yes) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = \frac{4}{9},$$

$$Gini(Own_{Condo} = no) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375\,,$$

$$Gini(split) = \frac{6}{10}\left(\frac{4}{9}\right) + \frac{4}{10}(0.375) = \frac{5}{12}$$

$$Gini(Unsecured_{Loan} = yes) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48\,,$$

$$Gini(Unsecured_{Loan} = no) = 1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0.32\,,$$

$$Gini(split) = \frac{5}{10}(0.48) + \frac{5}{10}(0.32) = 0.4$$

$$\boldsymbol{Gain(Annual_{income}) = 0.42 - 0.175 = 0.245}$$

$$Gain(Own_{Condo}) = 0.42 - \frac{5}{12} = 0.00333$$

$$Gain(Unsecured_{Loan}) = 0.42 - 0.4 = 0.02$$

c) Firstly, calculate the new Gini Index for Own_Condo and Unsecured_Loan.

$$Gini(Own_{Condo} = yes) = 1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0.32,$$

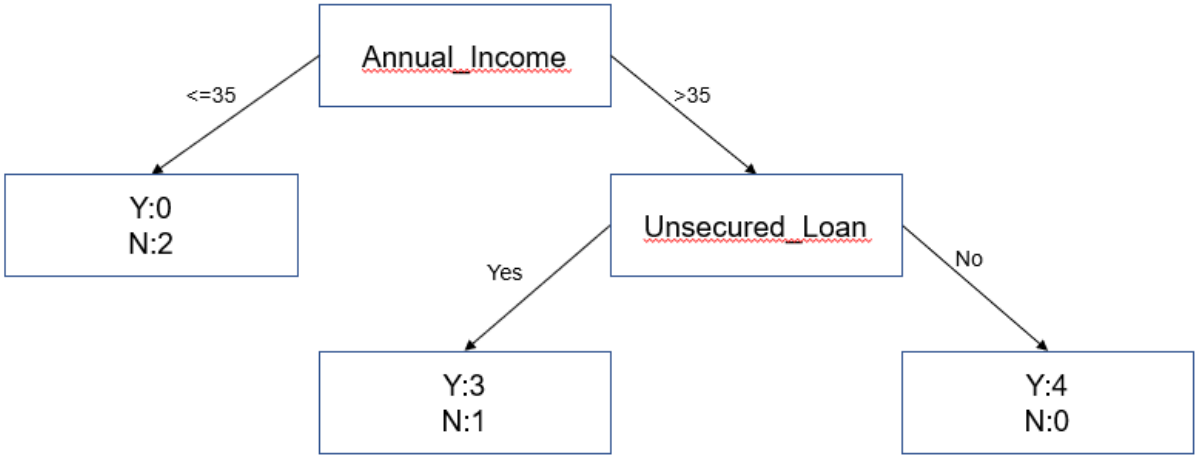
$$Gini(Own_{Condo} = no) = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0\,,$$

$$Gini(split) = \frac{5}{8}(0.32) + \frac{3}{8}(0) = 0.2$$

$$Gini(Unsecured_{Loan} = yes) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375\,,$$

$$Gini(Unsecured_{Loan} = no) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0\,,$$

$$\boldsymbol{Gini(split) = \frac{4}{8}(0.375) + \frac{4}{8}(0) = 0.1875}$$



d) Customer ID 011: Yes, Annual_Income > 35 → Unsecured_Loan = No.
Customer ID 012: No, Annual_Income <= 35.

3)

a)

$$MinMax\ Normalization: V' = \frac{V - min}{max - min}(new\_\max - new\_\min) + new\_min$$

$$V'_{\mathrm{A}} = \frac{V - 0}{10 - 0}(1 - 0) + 0 = \frac{V}{10}$$

$$V'_{\mathrm{B}} = \frac{V - 0}{5 - 0}(1 - 0) + 0 = \frac{V}{5}$$

Variable A	Variable B
0	0.2
0.2	0
0.3	0.4
0.4	0.6
0.5	0.6
0.8	1
1	0.8

	S1	S2	S3	S4	S5	S6	S7
S1		0.4	0.5	0.8	0.9	1.6	1.6
S2			0.5	0.8	0.9	1.6	1.6
S3				0.3	0.4	1.1	1.1
S4					0.1	0.8	0.8
S5						0.7	0.7
S6							0.4
S7							

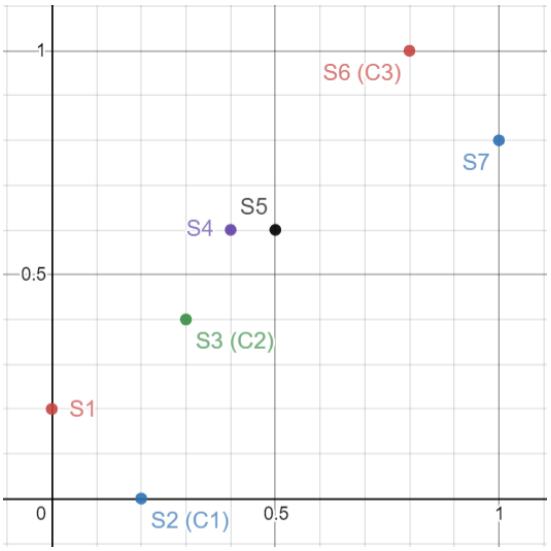
b) Normalize the new sample: [0.4,0.4].

	S1	S2	S3	S4	S5	S6	S7
New Sample	0.6	0.6	0.1	0.2	0.3	1	1

When K=1, you get S3 => benign.

When K=3, you get S3, S4, S5 => benign, malignant, malignant => malignant.

c)



	S1	S2	S3	S4	S5	S6	S7
S2	**0.4**			0.8	0.9		1.6
S3	0.5			**0.3**	**0.4**		1.1
S6	1.6			0.8	0.7		**0.4**

New Clusters = C1: {S1, S2}, C2: {S3, S4, S5}, C3: {S6, S7}

$$C1\ Centroid: \left(\frac{0+0.2}{2}, \frac{0.2+0}{2}\right) = (0.1, 0.1)$$

$$C2\ Centroid: \left(\frac{0.3+0.5}{2}, \frac{0.4+0.6}{2}\right) = (0.4, 0.5)$$

$$C3\ Centroid: \left(\frac{0.8+1}{2}, \frac{1+0.8}{2}\right) = (0.9, 0.9)$$

d) Single-link agglomerative hierarchical clustering = Min agglomerative hierarchical clustering

	S1	S2	S3	S4	S5	S6	S7
S1		0.4	0.5	0.8	0.9	1.6	1.6
S2			0.5	0.8	0.9	1.6	1.6
S3				0.3	0.4	1.1	1.1
S4					**0.1**	0.8	0.8
S5						0.7	0.7
S6							0.4
S7							

Join S4, S5

	S1	S2	S3	S4&S5	S6	S7
S1		0.4	0.5	0.8	1.6	1.6
S2			0.5	0.8	1.6	1.6
S3				**0.3**	0.7	0.7
S4&S5					0.8	0.8
S6						0.4
S7						

Join S3, S4 & S5

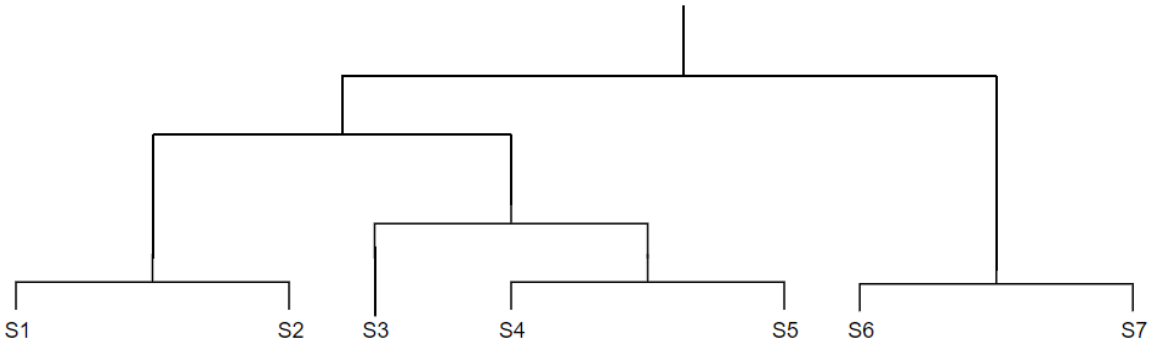
	S1	S2	S3&S4&S5	S6	S7
S1		**0.4**	0.5	1.6	1.6
S2			0.5	1.6	1.6
S3&S4&S5				0.7	0.7
S6					0.4
S7					

Join S1, S2

	S1&S2	S3&S4&S5	S6	S7
S1&S2		0.5	1.6	1.6
S3&S4&S5			0.7	0.7
S6				**0.4**
S7				

Join S6, S7

	S1&S2	S3&S4&S5	S6&S7
S1&S2		**0.5**	1.6
S3&S4&S5			0.7
S6&S7			



4)

a)

i) $\min\_sup = \frac{7}{2} = 3.5$

A	5
B	5
C	5
D	2
E	1
F	4
G	1
H	2
I	2

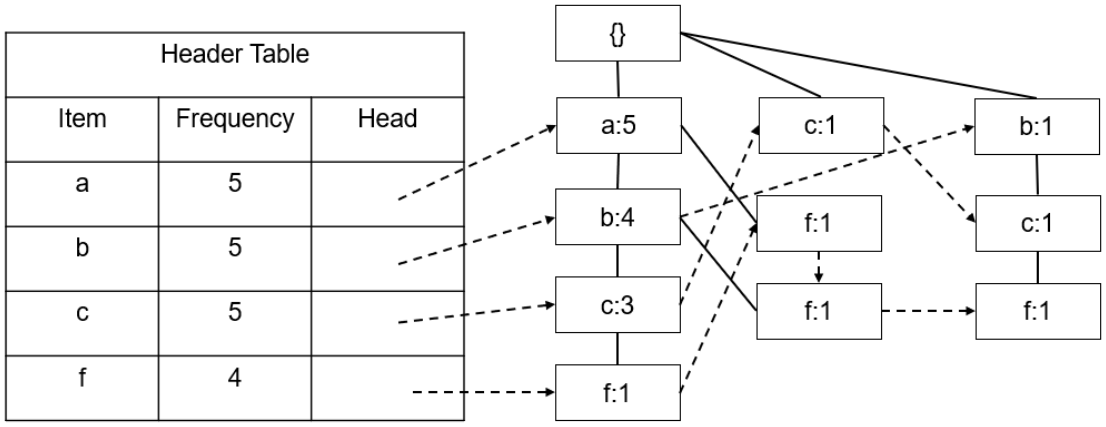
AB	4
AC	3
AF	3
BC	4
BF	3
CF	2

No need to iterate to 3rd level as AC is not frequent.
Frequent itemset = {a, b, c, f, ab, bc}

ii) A→B, confidence = 4/5
B→A, confidence = 4/5
B→C, confidence = 4/5
C→B, confidence = 4/5

iii) From (a), Ordered frequent itemset = {a, b, c, f}

Transaction ID	Items Purchased
0001	{a, b, c}
0002	{c}
0003	{a, b, c, f}
0004	{a, f}
0005	{a, b, c}
0006	{a, b, f}
0007	{b, c, f}



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) Graph-based clustering algorithm helps to reduce the amount of data that needs to be processed drastically due to sparsification. This helps to eliminate more than 99% of the entries in a proximity matrix, drastically reduce the amount of time required to cluster data and the size of the problems which can be handled is increased.
Jarvis-Patrick Clustering – Shared Nearest Neighbor Similarity Algorithm.

--End of Answers--