1. Provide brief answers (in about 50 words): (2 X 2 marks)
   a) List considerations to be made while deploying a data mining model.
   b) Compare cosine similarity and proximity-based similarity.

2. You work in the Technology Division of a major hospital. The hospital management do not seem to fully understand value of the data. Prepare a detailed illustrated note explaining to the management how the hospital will benefit from mining their data. (7 marks)

3. Answer the following with the supporting computations: (3 X 3 marks)
   a. Given the following three objects with five binary attributes.

	Attribute1	Attribute2	Attribute3	Attribute4	Attribute5
Object1	1	0	1	0	0
Object2	1	1	1	0	1
Object3	0	0	1	0	1

Among the three objects, find the most distinct pair of objects assuming attributes are symmetric binary variables? Which pair would be most distinct if attributes are asymmetric binary variables?

   b. Consider the confusion matrices of two classifiers.

Classifier1	PREDICTED CLASS		
ACTUAL CLASS		Class =Yes	Class =No
	Class= Yes	41	9
	Class= No	60	90

Classifier2	PREDICTED CLASS		
ACTUAL CLASS		Class =Yes	Class =No
	Class= Yes	40	10
	Class= No	50	100

Using the following cost of error matrix, identify which one is better.

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	-1	10
	Class=No	1	0

   c. A survey was done on sample of IT employees about their salaries. Given below are salaries of hardware and software employees (in thousands of rupees).

      Hardware employees: 30, 25, 45, 50, 40, 34

      Software employees: 20, 35, 40, 44, 30, 55, 60, 25, 51, 70, 65

   Make a quantile-quantile plot for salaries of these two groups.

4. Answer with brief justification
    a. A rule-based classifier is built through indirect method from a decision tree. The decision tree turns out to be overfit on training data. What will be the impact on rule-based classifier? (3 marks)
    b. The following training dataset provides income (can be high-H, medium-M, or low-L) and education level (can be high-H or low-L) of customers who bought a smart watch. We intend to construct decision tree to predict who buys smart watch. Identify the attribute for root node using the Gain Ratio. (5 marks)

Income	H	H	L	M	M	M	L	H	H	M	H	L	L	M
Education	H	H	H	H	L	L	L	H	L	L	L	H	L	H
Bought Smart watch	N	Y	Y	Y	Y	N	Y	N	Y	Y	Y	Y	Y	N

5. Answer the following:
    a. Through some initial market study (captured in the contingency table below), a marketing manager started promoting skimmed milk powder to the coffee buyers. Do you agree with him? You answer should be justified with association mining metrics. Assume thresholds for support = 30%, confidence = 60%. [4 Marks]

Bought	**Coffee**	**No Coffee**
Skimmed Milk	1000	875
No Skimmed Milk	500	125

    b. When do you prefer apriori algorithm over FP-tree for doing frequent pattern mining? (3 marks)

6. Answer the following:
    a. We have the following distance matrix among 6 objects. Perform agglomerative hierarchical clustering with MIN approach and draw dendrogram. Show the intermediate steps. (4 marks)

	P1	P2	P3	P4	P5	P6
P1	0	0.155	0.101	0.077	0.119	0.264
P2	0.155	0	0.098	0.204	0.274	0.254
P3	0.101	0.098	0	0.174	0.204	0.309
P4	0.077	0.204	0.174	0	0.113	0.222
P5	0.119	0.274	0.204	0.113	0	0.334
P6	0.264	0.254	0.309	0.222	0.334	0

    b. You have been given a dataset of 800 objects. It is known that there are 8 natural clusters. You plan to perform K-means clustering starting with 8 random initial seeds? What is the probability that initial seeds come from distinct clusters? Comment on the results and consequences. [2+2 marks]

7. Answer the following:
    a. You are evaluating a credit card fraud detection system that has 98% accuracy. It is known that typically there are 10 fraudulent transactions per million. Assuming that the system has symmetric accuracy for both classes (Fraud, Genuine) of transactions, estimate false alarm rate. (4M)
    b. When do we prefer supervised and unsupervised approaches for sentiment analysis? (3M)