Question 1:

1. (a) Describe three challenges to data mining regarding data mining methodology and user interaction issues.

1. (b) Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):
(a) Compute the Euclidean distance between the two objects.
(b) Compute the Manhattan distance between the two objects.
(c) Compute the Minkowski distance between the two objects, using q = 3.

1. (b')
Suppose that for two vectors *A* and *B*, we know that their Euclidean distance is less than *d*. What can be said about their Manhattan distance?

2. (a) Give a short example to show that items in a strong association rule actually may be negatively correlated.
(b) The following contingency table summarizes supermarket transaction data, where hot dogs refers to the transactions containing hot dogs, $\overline{hot\ dogs}$ refers to the transactions that do not contain hot dogs, hamburgers refers to the transactions containing hamburgers, and $\overline{hamburgers}$ refers to the transactions that do not contain hamburgers.

	hot dogs	$\overline{hot\ dogs}$	Σ_{row}
hamburgers	2000	500	2500
$\overline{hamburgers}$	1000	1500	2500
Σ_{col}	3000	2000	5000

(a) Suppose that the association rule "hot dogs ⇒ hamburgers" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?
(b) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers? If not, what kind of correlation relationship exists between the two?

2. (b) A database has five transactions. Let min sup = 55% and min conf = 75%.

TID	*items_bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

(a) Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.
(b)) List all the strong association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and itemi denotes variables representing items (e.g., "A," "B,"):

$$\forall x \in transaction,\ buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3) \quad [s, c]$$

2.(b') Explain FP-Tree algorithm using an example.

3. (a) Cluster the following eight points (with (x, y) representing locations) into three clusters:

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-

P(a, b) = |x2 – x1| + |y2 – y1|

Use K-Means Algorithm to find the three cluster centers after the second iteration.

3 (b). For a given k=2, cluster the following data set using K mediod clustering algorithm.

Point	x-axis	y-axis
1	7	6
2	2	6
3	3	8
4	8	5
5	7	4
6	4	7
7	6	2
8	7	3
9	6	4
10	3	4

3 (b'). Explain hierarchical clustering algorithm using an example.

4 (a). Explain non-linear SVM in detail.

4 (b). Illustrate using an example how dimensionality gets reduced in PCA.

4(b’). For the table given below:

Attributes are Color , Type , Origin, and the subject, stolen can be either yes or no.

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Classify a Red Domestic SUV using Naive Bayesian Classifier.

5 (a). The following table consists of training data from an employee database. The data have been generalized. For example, “31 ... 35” for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row.

department	*status*	*age*	*salary*	*count*
sales	senior	31...35	46K...50K	30
sales	junior	26...30	26K...30K	40
sales	junior	31...35	31K...35K	40
systems	junior	21...25	46K...50K	20
systems	senior	31...35	66K...70K	5
systems	junior	26...30	46K...50K	3
systems	senior	41...45	66K...70K	3
marketing	senior	36...40	46K...50K	10
marketing	junior	31...35	41K...45K	4
secretary	senior	46...50	36K...40K	4
secretary	junior	26...30	26K...30K	6

Let status be the class label attribute.

(a) How would you modify the basic decision tree algorithm to take into consideration the count of each generalized data tuple (i.e., of each row entry)?

(b) Use your algorithm to construct a decision tree from the given data

5 (b). Using an example, describe in detail Random Forest algorithm.

5(b’). (i) It is important to calculate the worst-case computational complexity of the decision tree algorithm. Given data set, D, the number of attributes, n, and the number

of training tuples, |D|, show that the computational cost of growing a tree is at most n × |D| × log(|D|).

(ii) Why is naive Bayesian classification called “na¨ıve”? Briefly outline the major ideas of na¨ıve Bayesian classification.