Data Analytics I: Assignment 1

1) Given the following frequent itemsets what candidates will Apriori compute for the next database scan?
(i) AB, AC, AD, BC, BD, CD, AE

(A) ABC, ABD, ACD, BCD, ABE, ACE, ADE, BCD, ABCD
(B) ABC, ABD, ACD, BCD, ABE, ACE, ADE, BCD
(C) ABC, ABD, ACD, BCD, ABCD
(D) ABC, ABD, ACD, BCD
(E) Null-set

(ii) ABC, ABD, ACD, BCD, BCE, CDE

(A) ABCD, BCDE, ACDE, ABCDE
(B) ABCD, BCDE, ACDE
(C) ABCD, BCDE
(D) ABCD
(E) Null-set

2) Use naive bayes on the following data to classify Red Domestic SUV.

Example#	Colour	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

(i) What is P(Red Domestic SUV / Stolen) as computed in naive bayes? _____

(ii) What is P(Red Domestic SUV / Not stolen) as computed in naive bayes? _____

3) For the data in Q2, if we run ID3, what is the information gain of each attribute in the first level?

(A) Example#: _____
(B) Colour: _____
(C) Type: _____
(D) Origin: _____

(E) Entropy at level 1 is: _____

4) Data: {(Ram,64,60),(Shyam,60,61),(Gita,59,70),(Mohan,68,71)}. Run 2 iterations of k-means algorithm using euclidean distance and k=2. Choose Shyam and Gita as initial means.

(i) The clusters after 2 iterations are: _____ and _____
(ii) The clustering quality is: _____
