Unique Paper Code	**: 32347611**
Name of the Paper	**: Data Mining**
Name of the Course	**: B.Sc. (H) Computer Science**
Semester	**: Semester -VI**
Duration of Examination	**: Two Hours**
Maximum Marks	**: 75 Marks**
Year of Admission	**: 2015, 2016, 2017**

Instructions for Candidates

Attempt Any Four questions. All Questions carry equal marks.

Q 1. Given the following table, classify all the attributes appearing in the table as *binary, discrete or continuous*. Also classify them as qualitative (*nominal or ordinal*) or quantitative (r*atio or interval*). Justify your answer in each case. Show the normalization (scaling data between 0 and 1) of values in the age attribute column. How can you handle missing values in Age column and Height column? Replace Y and N respectively by 1 and 0 in the first column. Replace M and F respectively by 0 and 1 in the second column. You will get one binay vector each for Smoker attribute and Gender attribute. Find out the similarity measure between these two vectors using Jaccard coefficient.

Smoker	Gender	Age	Height	Marital Status
Y	F	32	Tall	Married
Y	M	34	Medium	Marries
N	F	39	Medium	Single
Y	M	41	Tall	Single
Y	M	25	Tall	Divorcee
N	M	36	Tall	Single
Y	F	45	Short	Married
Y	M	31	Tall	Single
N	M	29	Medium	Divorcee
N	F	51	Tall	Single
Y	F	38	Short	Married

Q 2. Given the following binary classification problem :

Instance	A1	A2	Target Class
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-
7	F	F	-
8	T	F	+
9	F	T	-
10	T	F	-

Calculate separately the information gain when splitting is done on A1 and on A2. Which attribute would the decision tree induction algorithm choose? Calculate separately the gain in the Gini index when splitting is done on A1 and A2. Which attribute would the decision tree induction algorithm choose? Is it possible that information gain and the gain in Gini index favour different attributes? Explain your answer.

Q 3. Consider the one dimensional labeled data set given below:

X:	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
Y:	-	-	+	+	-	-	+	+	-	-

Classify the data point x= 5.0 according to its 3- and 5- nearest neighbour using majority vote.

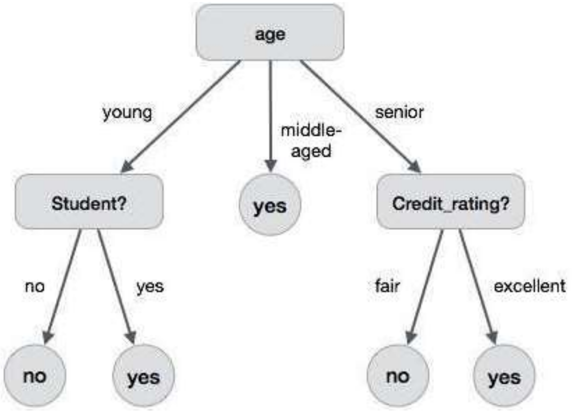
Suppose that there are a total of 60 data mining related documents in a library of 200 documents. Suppose that a search engine retrieves 20 documents after a user enters 'data mining' as a query, of which 5 are data mining related documents. What are the precision and recall?

Q 4. Consider the Market Basket dataset shown below:

Transaction ID	**Items Bought**
0001	{a,d,e}
0024	{a,b,c,e}
0012	{a,b,d,e}
0031	{a,c,d,e}
0015	{b,c,e}
0022	{b,d,e}
0029	{c,d}
0040	{a,b,c}
0033	{a,d,e}
0038	{a,b,e}

Compute support for item sets {e}, {b,d}, {b,d,e}, {a,b,c,e} and {a,b,c,d,e}. Find all association rules that can be generated from item set {b,d,e} along with the confidence of each rule. Is confidence a symmetric measure?

Q 5. Generate all the rules given the following decision tree:



Arrange the generated rule according to class based ordering and classify the following tuples:

Age = young	student = no	loan_approval=?
Age = senior	credit_rating = excellent	loan_approval=?

Consider a training set that contains 90 positive examples and 300 negative examples. For each of the following rules:

R1: A----->+ (Covers 30 positive and 10 negative examples)

R2:B---->+ (Covers 90 positive and 80 negative examples)

R3:C ---->+ (Covers 4 positive and 1 negative example)

Determine which is the best and which is the worst candidate rule according to rule Accuracy.

Q 6. Use K-means algorithm and Euclidean distance to cluster five data points (A4-A8) given below, into 3 clusters. The coordinates of the data points are:

A1(2,8), A2(2,5), A3(1,2), A4(5,8), A5(7,3), A6(6,4), A7(8,4), A8(4,7).

Use A1, A2, A3 as initial centroids. For which situations K-mean clustering will give good results and when will it fail to produce good results?