BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
Work Integrated Learning Programmes (WILP) Division
M. Tech -Data Science and Engineering
I Semester 2019-20
End-Semester Examination

Course No	:	**DSECF ZC415**
Course Name	:	**Data Mining**
Nature of Exam	:	Open Book
Max Marks	:	50
Duration	:	2-Hour
Date of Exam	:	(AN)

No. of page: 3
No. of questions: 4

Note:
1. Please read and follow all the instructions given on the cover page of the answer booklet.
2. **Start each answer from a fresh page. All parts of a question should be answered consecutively**.
3. Please ensure that your answers cover necessary technical details, avoiding unnecessary text and diagrams.

Question 1

a) Ramesh is an investor. His portfolio primarily tracks the performance of the Nifty and Ramesh wants to add the stock of ABC Corp. Before adding the stock to his portfolio, he wants to assess the directional relationship between the stock and the Nifty.
Ramesh does not want to increase the unsystematic risk of his portfolio. Thus, he is interested in owning securities in the portfolio that tend to move in the same direction.
Considering Ramesh's criteria and using the data set given below, provide a recommendation whether Ramesh should invest in ABC Corp. stock? **[5 + 1 = 6 marks]**

Year	**2013**	**2014**	**2015**	**2016**	**2017**
Nifty	1692	1978	1884	2151	2519
ABC Corp	68	102	110	112	154

b) A data scientist has one-dimensional data and she is checking if there are outliers in the data. Compare pros and cons of using statistical approach of ±3-sigma dispersion and boxplot approach with 1.5*IQR . **[4 marks]**

c) Apply equi-width and equi-depth binning method on the following dataset to create sets of 3 bins.
[23, 8, 2, 20, 11, 1, 29, 30, 21]
[4 marks]

Question 2

a) Suppose you have collected a set of 1,000,000 labeled data points and you build a decision tree classifier from them. You then choose 100 of these points at random, and find that your classifier returns the correct answer on all of them. Can you conclude that your algorithm works with a 0% error rate on any input? Why or why not? **[3 marks]**

b) An FMCG company prepared a training set that contains one hundred records for its new ayurvedic toothpaste (T) and four hundred records from its competitors offerings of the same type of product (C). The following classification rules are built from this training set, where P, Q and R denote the subsets of the attribute values in the records which consumers might consider to buy a toothpaste. Using FOIL gain metric find out, which rule would the company be interested in for its predictive modelling? **[6 Marks]**

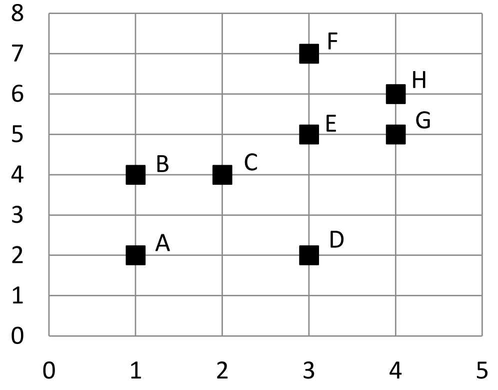
R0: ϕ → T (the initial NULL rule that covers one hundred T and four hundred C records)
R1: P → T (covers four T and one C records)
R2: Q → T (covers thirty T and ten C records)
R3: R → T (covers one hundred T and ninety C records)

c) Through some initial market study (captured in the contingency table below), a marketing manager believed the coffee buyers tend to buy skimmed milk powder. Do you agree with him? Your answer should be mathematically justified taking support, confidence and lift measures. Assume thresholds for support = 30%, confidence = 60%. **[4 marks]**

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

Question 3

Answer the following questions in the context of shown figure of 8-point two-dimensional dataset (A to H): **[5x2 = 10 Marks]**



a) BIRCH method is run on the given dataset. Assume that at some point of time the points {A, B, C, D} are placed in one cluster. What will be the Clustering Feature and Radius of this cluster?

b) DBSCAN method is run on the given dataset. Given that MinPts = 3 and ε = 2 units. Find out the Core, Border and Noise points. Manhattan distance needs to be taken as the distance measure and while counting the MinPts for a point, the point itself need to be included in the counting.

Question 4

a) Term frequency matrix for the five articles (A1 to A5) is shown below. Answer the following questions: **[3+7= 10 Marks]**

1) What is the TF-IDF value for (A4, Corona)?
2) Find the cosine similarity between articles? Identify the two articles that are the most similar.

Article/Terms	**Trump**	**JNU**	**AAP**	**Corona**	**Divestiture**
A1	14	1	0	6	3
A2	0	21	5	0	0
A3	0	15	18	0	5
A4	5	2	0	12	0
A5	0	0	5	0	10

b) Often TF-IDF is used as multidimensional data in classifying document collections. Can you think of a weakness in the approach and how you may overcome? **[3 marks]**