Roll Number:_____ No. of Pages: 03

National Institute of Technology, Hamirpur (H.P.)

Department of Computer Science and Engineering

M. Tech./Ph.D.: Semester- I (AI)	Course Code: CS-719 Course Name: Data Mining
December 17, 2020	Thursday, 15:00 – 17:00 Hrs
Time: 2 Hours, M. Marks: 50 Marks	Name Of Faculty: VKC

Note: Attempt all questions in proper sequence. Assume missing data, if any, suitably.

Q1 Consider an organization tested eighteen randomly chosen persons. The age and fat data are mentioned in the following table: 9 Marks

Age	23	23	27	27	39	41	47	49	50
%Fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
Age	52	54	54	56	57	58	58	60	61
%Fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

a) Compute the mean, median and standard deviation of age and %fat.

b) Draw the boxplots for Age and %Fat.

c) Normalize the two variables based on z-score normalization.

d) Compute the Pearson correlation coefficient. Are these two variables positively or negatively correlated?

Q2 Compute the hierarchical F-measure for the eight objects {p1, p2, p3, p4, p5, p6, p7, p8} and hierarchical clustering shown in **Fig. 1**. Class A contains points p1, p2, and p3, while p4, p5, p6, p7, and p8 belong to class B. 9 Marks

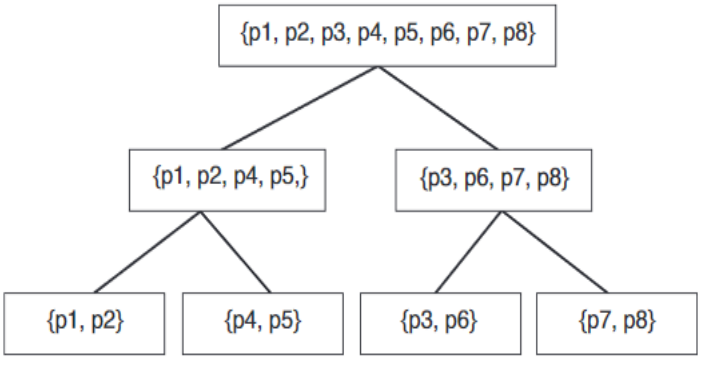

Fig. 1. Hierarchical Clustering

Q3 Consider the dataset consists of five transactions. Assume min_suppport = 60% and min_confidence= 80%. 8 Marks

TID	Items Bought
T1	{M,O,N,K,E,Y}
T2	{D,O,N,K,E,Y}
T3	{M,A,K,E}
T4	{M,U,C,K,Y}
T5	{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 of the strong association rules (with support ***s*** and confidence ***c***) matching the following meta rule, where X is a variable representing customers, and $item_i$ denotes variables representing items (e.g., "A", "B", etc.):

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

Q4 Consider a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. 8 Marks

(a) Enumerate three classes of schemas that are popularly used for modeling data warehouses.

(b) Draw a schema diagram for the above data warehouse using one of the schema classes listed in Question 4(a).

(c) Starting with the base cuboid [day; doctor; patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2020?

(d) To obtain the same list, write an SQL query assuming the data is stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge).

Q5 Consider the following four pages with damping factor = 0.85 and their links in context to the Page Rank algorithm. 8 Marks

Page A has page rank of 1 and has one link to B.
Page B has page rank of 2 and has two links to C and D.
Page C has Page rank of 3 and has two links to B and D.
Page D has page rank of 2 and has three links to A, B, and C.

a) Find page rank for all the web pages using page rank algorithm.

b) Which page has the highest page rank?

Q6 Consider the following data set for class problem. Calculate the gain in the Gini index when splitting on attributes ***X***, ***Y*** and ***Z*** having values T=True and F=False. 8 Marks

X	***Y***	***Z***	**Class**
T	T	T	I
F	F	F	II
T	T	F	III
T	F	T	I
F	T	F	III
F	F	F	II
F	F	T	I
T	F	F	II
F	T	F	III
T	T	F	III

a) Which attribute would the decision tree induction algorithm choose as root attribute in the decision tree?

b) Build the decision tree for class problem.