Dr Nagendra Pratap Singh 2/12/2022 E 88

National Institute of Technology Hamirpur

End Semester Examination
B.Tech. CSE, 7^{th} Semester
Data Analytics (CS-432)

Time: 02 : 30 PM–05 : 30 PM **Date:** 02/12/2022
Duration: 3.0 Hours **Max. Marks:** 50

Note: Answer all questions and each question having equal marks.

1. (a) Consider the following Minitab display of two data sets C_1 and C_2. **[05 marks]**

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
C1	20	20.00	1.62	7.26	7.00	15.00	20.00	25.00	31.00
C2	20	20.00	1.30	5.79	7.00	20.00	22.00	22.00	31.00

   i. What are the respective means and the respective ranges?
   ii. Which data set seems more symmetric and Why?
   iii. Compare the interquartile ranges of the two sets. How do the middle halves of the data sets compare?

   (b) The mean and standard deviation of a set of N_1 observations are $\bar{X}_1$ and S_1, respectively while the mean and standard deviation of another set of N_2 observations are $\bar{X}_2$ and S_2, respectively. Find the standard deviation of the combined set of (N_1+N_2) observations. **[05 marks]**

2. (a) Trace the results by using the Apriori algorithm on the grocery store example with support threshold S=33.33% and confidence threshold C=60%. Show the candidate and frequent itemsets for each database scan. Enumerate all the final frequent itemsets. Also indicate the association rules that are generated and highlight the strong ones, sort them by confidence. **[08 marks]**

Transaction ID	Items
T_1	H, B, K
T_2	H, B
T_3	H, C, P
T_4	P, C
T_5	P, K
T_6	H, C, P

   (b) What is the difference between Existing Apriori algorithm and an Improved Apriori algorithms proposed by author 'Mohammed Al-Maolegi et al'. **[02 marks]**

3. Answer the following questions **[10 marks]**

   (a) What is the difference between Quantitative and Qualitative data?

   (b) Explain the 3V's characteristic of Big Data.

   (c) If the mean of 100 observations is 50 and their standard deviation is 5 then what is the value of sum of all squares of all the observations.

   (d) What are the different data objects in R?

ID	Fever	Breathing issues	Cough	Infected
1	YES	YES	YES	YES
2	No	No	No	No
3	YES	No	YES	No
4	YES	YES	No	YES
5	YES	YES	YES	YES
6	YES	YES	No	YES
7	No	No	YES	No
8	YES	YES	No	YES
9	YES	No	YES	No
10	No	YES	YES	No
11	No	YES	YES	YES
12	YES	No	YES	YES
13	No	No	YES	No
14	No	YES	YES	YES

4. (a) Construct a decision tree for the following data set using information gain as the attribute selection criteria (ID3 Algorithm). **[08 marks]**

   (b) What do you understand by Hadoop Map Reduce? Explain Map and Reduce task?

   [02 marks]

5. The following data indicate the number of tablets produced every day (in thousands) by five separate technicians utilizing four distinct machine types. **[10 marks]**

Workers	A	B	C	D
P	54	48	57	46
Q	56	50	62	53
R	44	46	54	42
S	53	48	56	44
T	48	52	59	48

Using two-way ANOVA test:

(a) Test whether the mean productivity of the different machines is the same.

(b) Test whether the 5 technicians differ with respect to the mean productivity.