Q1: Imagine that students in a class are friends on a social media platform. The instructor wants send a notice to these students. So, the instructor sends a message to the student who has many other students in the class thinking that the message can reach to as many students as possible. In this scenario, what type of data would you use and how will you create the dataset?

Q2: An organization is screening job candidate applications for various positions. The dataset consist of height (in CM) and Weight (in Kilos) of applicants. An applicant is considered suitable if his/her BMI is between 21-24.

1. What concepts learnt from the data preprocessing you will use?
2. How will you convert the given continuous values into a binary feature meaningful to the application?

Q3. Consider the problem of finding the K nearest neighbors of a data object. Use the below given object dataset for answering the questions.

Algorithm: Finding the K nearest neighbors

For i = 1 to number of data objects

Find the proximity of the ith object to all other objects

Sort these proximities in decreasing order

(Keep track of which object is associated with each proximity value)

Return the first K objects of the sorted list

End for

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data object | Name | Gender | Salary | Age |
| Obj1 | AFD | Male | High | Old |
| Obj2 | SRE | Male | Low | Middle |
| Obj3 | SDF | Female | High | Young |
| Obj4 | AQW | Male | Low | Middle |
| Obj5 | ASW | Female | High | Young |

1. Which are the 3 nearest object of Obj1 using
   1. Euclidian distance measure
   2. Cosine similarity measure
2. Do you see any challenges with respect to this algorithm which are associated with data quality that might impact the outcome? How you will fix it?

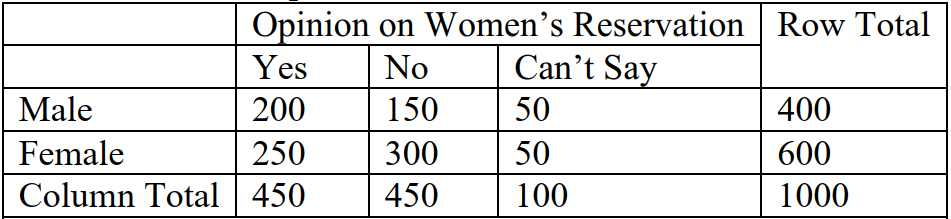
Q4. Consider the training examples shown in below table for a binary classification problem.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Customer ID | Gender | Car Type | Shirt Size | Class |
| 1 | M | Family | Small | C0 |
| 2 | M | Sports | Medium | C0 |
| 3 | M | Sports | Medium | C0 |
| 4 | M | Sports | Large | C0 |
| 5 | M | Sports | Extra Large | C0 |
| 6 | M | Sports | Extra Large | C0 |
| 7 | F | Sports | Small | C0 |
| 8 | F | Sports | Small | C0 |
| 9 | F | Sports | Medium | C0 |
| 10 | F | Luxury | Large | C0 |
| 11 | M | Family | Large | C1 |
| 12 | M | Family | Extra Large | C1 |
| 13 | M | Family | Medium | C1 |
| 14 | M | Luxury | Extra large | C1 |
| 15 | F | Luxury | Small | C1 |
| 16 | F | Luxury | Small | C1 |
| 17 | F | Luxury | Medium | C1 |
| 18 | F | Luxury | Medium | C1 |
| 19 | F | Luxury | Medium | C1 |
| 20 | F | Luxury | Large | C1 |

1. Compute the Gini index for the overall collection of training examples.
2. Compute the Gini index for the Gender attribute.
3. Which attribute is better Gender, Car Type or Shirt size?

Q5: A sample of 1000 students have been surveyed on Women’s reservation. The respondents have been classified using gender (male or female) and by opinion (reservation for women, No

reservation, or No opinion). Table shows the result of the survey:



You are now asked to find whether the men's opinion on women's reservation differ significantly

from that of women using chi2 test with 0.05 level of significance. From the standard chi2 test

table the P-value for the chi square value you compute is 0.0003. Answer questions i-iv.

i. State your hypothesis.

ii. Compute the expected frequency (under the null hypothesis)

iii. Select the test statistic

iv. Make your conclusion based on the value.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*