Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division

Comprehensive Examination (Regular)

Course No.	: PCAM* ZC111
Course Title	: FEATURE ENGINEERING
Nature of Exam	: Open Book
Weightage	: 40%
Duration	: 2 Hours
Date of Exam	:

No. of Pages =**3**
No. of Questions = **4**

Note:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q1. See the following visual, which shows year-over-year change ("YoY," measured as percent change in dollar sales volume) for food brands from a pet food manufacturer.
[1 +1 + 1 + 1 + 1 + 5 = 10]

a) You need to present this data in live interaction and want to begin by talking about the Lifestyle brand line: Lifestyle, Diet Lifestyle, and Lifestyle Plus. How would you visually indicate to your audience to look at those points of data?
b) Now you want to talk about the Feline brand group, (includes all of the brands with "Feline" in their name). The branding for this line of food has a red colored logo. How would you indicate to your audience that they should focus here?
c) Next, need to discuss the brands that had year-over-year declines. How could you draw your audience's attention there?
d) Now imagine that within the declining brands, you need to talk specifically about the two brands that declined the most. How might you achieve this?
e) Let's assume that you want to talk about the brands that had year-over-year increases in sales. How would you draw your audience's attention there? Is there any similarity in a way that you used to direct attention to the decreasing brands?
f) Now you need to prepare a final comprehensive representation to be shared that highlights each of the takeaways outlined previously: Lifestyle brands, Feline brands, decreasing brands (differentiating those decreasing most), and increasing brands (highlighting those that increased most). How would you achieve this? How would you pair this with explanatory text and make it clear how the text relates to the data?

Q2. The CoWax application receives several requests from the Indian citizens for registering their details and slot booking for the vaccination. **[3 + 2 + 3 + 1 + 1 = 10]**
This data is captured in database of the application as listed below format.

- User ID (Unique identification for each user of application)
- Aadhar ID of user who has registered
- Mobile number
- Name
- Secret Code
- Year of Birth
- Photo ID (3 types : Aadhar, PAN, Passport)
- ID Number
- Type of vaccine (3 types: Covishield, Covaxin, Sputnik)
- Gender (Male, Female)
- Marital Status (Family, Single, Other)
- Age Group (18 to 44 , 45 to 60, above 60)
- Number of doses
- Service Type (Free, Paid)
- Charges Per dose
- Date of registration
- Date of appointment booking
- Date of first dose
- Date of second dose

a) Draw three appropriate charts for one categorical, numeric and binary attribute. What insights can be obtained based on these charts?
b) Identify the redundant attributes present in the dataset?

c) If you are asked to develop following three predictive models based on the above mentioned predictor attributes, then what will be the potential target attributes suitable for model building?
   - Multiple linear regression
   - Logistic regression
   - Decision tree

d) Draw simple dataset schema that will be considered for the development of Multiple Linear regression model along with sample record in it.

e) Identify two potential issues that you foresee while dealing with numeric attributes given in the dataset.

Q3. Consider the problem of finding the K nearest neighbors of a data object. Use the below given object dataset for answering the questions**.** **[8 + 2 = 10]**

Algorithm: Finding the K nearest neighbors

For i = 1 to number of data objects
Find the proximity of the i^{th} 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

a) Which are the 3 nearest object of Obj1 using
   I. Euclidian distance measure
   II. Cosine similarity measure

b) 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.
[0.5 + 2 + 1.5 + 2 + 2.5 + 0.5 + 1 = 10]

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

a) Compute the Gini index for the overall collection of training examples.
b) Compute the Gini index for the Customer ID attribute.
c) Compute the Gini index for the Gender attribute.
d) Compute the Gini index for the Car Type attribute.
e) Compute the Gini index for the Shirt Size attribute.
f) Which attribute is better Gender, Car Type or Shirt size?
g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.

****