**Birla Institute of Technology & Science, Pilani**

**Work-Integrated Learning Programmes Division**

**Second Semester 2019-2020**

**Comprehensive Examination (Regular)**

Course No. : PCAM\* ZC111

Course Title : FEATURE ENGINEERING

Nature of Exam : Closed Book

No. of Pages =**3** No. of Questions =5

Weightage : 30%

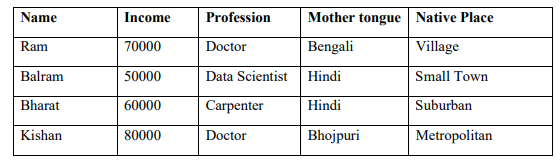
Duration : 2 Hours

Date of Exam : ~~Friday, 20/09/2019 (AN)~~

Note:

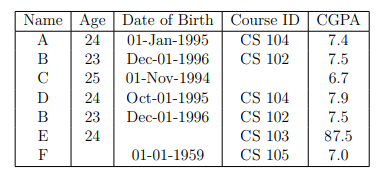
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q1. Based on the information given in the table below, find most similar and most dissimilar persons among them. Apply min-max normalization on income to obtain [0,1] range. Consider profession and mother tongue as nominal. Consider native place as ordinal variable with ranking order of [Village, Small Town, Suburban, Metropolitan]. Give equal weight to each attribute. **[10]**



Q2. List four important things that you should know before creating a pie chart. **[4]**

Q3. Consider the given data set for carrying out some data science activity. **[4 + 2 = 6]**



1. Identify and correct four potential issues with this dataset.
2. List down two major objectives behind data preprocessing step of CRISP-DM framework.

Q4.(a) Consider the following training data set (with two attributes and two possible classes Y, N) for a binary class problem. The attributes are nominal with two possible values. We intend to create decision tree model. **[3 + 1=4]**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Income | H | H | H | H | H | L | L | L | H | H |
| Education | L | H | H | L | H | L | L | L | H | L |
| Will Buy | N | N | N | Y | N | Y | Y | Y | Y | Y |

1. Calculate the gain in the Gini index when splitting on *Income* and *Education*.
2. Which attribute would the decision tree induction algorithm choose?

Q4. (b) For each of the following meetings, explain which phase in the CRISP-DM process is presented :-  **[ 4]**

1. Project manager meets analysts to discuss how useful and accurate DM model is.
2. Manager meets data warehouse stakeholders to discuss how data will be collected.
3. Data mining consultant meets the VP of marketing, who says that he is interested to see how data mining can be used in marketing campaigns.
4. The analysts meet to discuss whether decision tree or SVM should be used.

Q4. (c) The Data mining process is autonomous which does not require human oversight. Justify or Invalidate. **[2]**

Q5. As a data analyst at a national clothing retailer, you are gearing up for this year’s back-to-school shopping season. You’ve analyzed survey data from last year’s back-to-school shopping to understand customers’ experience—what they liked and what they didn’t like. You believe the data reveals some clear opportunities and want to use it to inform the strategy for this year’s back-to-school shopping season across your company’s stores. What is the story here? How would you visualize the data in following figure to lend insight into what we should focus on in this situation? Make assumptions about the scenario as needed. **[10]**

