Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
First Semester 2019-2020

Comprehensive Examination (Regular)

Course No.	: PCAM* ZC111
Course Title	: FEATURE ENGINEERING
Nature of Exam	: Closed Book
Weightage	: 30%
Duration	: 2 Hours
Date of Exam	: ~~Friday, 20/09/2019 (AN)~~

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.

Q.1 Answer the following questions in short: **[2 * 2 = 4]**

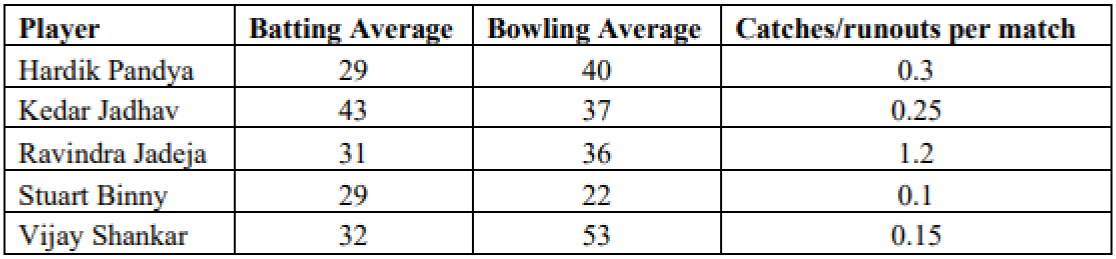
a) Raju measures the pressure of all tires coming into his garage and record the values. Unknown to him, his tire gauge is miscalibrated and adds 3 psi to each reading. Using the definition of noise used in the textbook, is this error introduced by the tire gauge considered noise? Answer “yes” or “no” and justify your answer in one line.

b) For each of the following meetings, explain which phase in the CRISP-DM process is represented:

i) Managers want to know by next week whether deployment will take place. Therefore, analysts meet to discuss how useful and accurate their model is.

ii) The data mining project manager meets with the data warehousing manager to discuss how the data will be collected

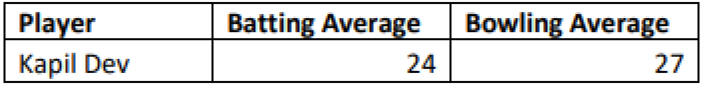
Q.2. You are the chief selector of the Indian cricket team and you are tasked with selecting the best all-rounder for the Indian world cup squad. Below is the list of all-rounders who are available for selection and their respective batting, bowling and fielding stats. ` **[2 + 4 + 4 =10]**



Player	Batting Average	Bowling Average	Catches/runouts per match
Hardik Pandya	29	40	0.3
Kedar Jadhav	43	37	0.25
Ravindra Jadeja	31	36	1.2
Stuart Binny	29	22	0.1
Vijay Shankar	32	53	0.15

a. Identify the player who is an outlier based on “Catches/runouts per match”. Explain, how you decided the outlier.

b. Below are Kapil Dev’s stats (fielding stats not available). Using Manhattan distance, find out which all-rounder is most “Kapil-like”.



Player	Batting Average	Bowling Average
Kapil Dev	24	27

c. Do you think “Cosine similarity” would give same result as “Manhattan Distance” in part 2)?

Q.3. Consider following taxpayers dataset. Several interesting characteristics of taxpayer like marital status, DOB, income, refund status along with tax evasion status are captured. Using this dataset, answer the following sub questions. **[2 + 6 + 2 = 10]**

Txn_ID	Marital_Status	Date_of_Birth	Taxable_Income	Refund_Status	Evasion_Status
1	Single	20 March 82	125K	Yes	No
2	Married	31 July 86	100K	No	No
3	Single	17 Jan 89	70K	No	No
4	Married	25 Aug 84	120K	Yes	No
5	Divorced	17 Sept 91	95K	No	Yes
6	Married	2 Nov 89	60K	No	No
7	Divorced	8 Nov 87	220K	Yes	No
8	Single	9 Feb 81	85K	No	Yes
9	Married	18 Apr 85	75K	No	No
10	Single	7 March 87	90K	No	Yes

You need to use this dataset to predict the probability that a taxpayer will evade the tax.

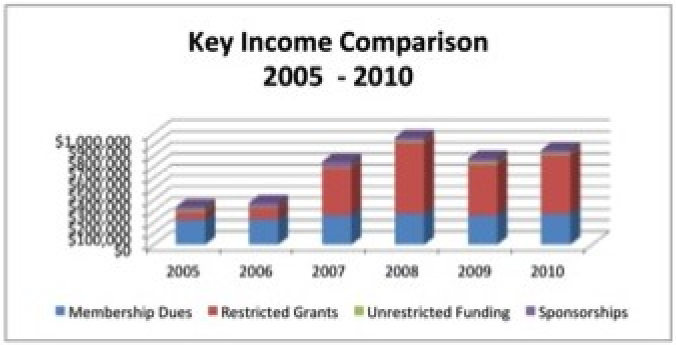
a. What is the modelling technique that will be useful for above requirement? Why?

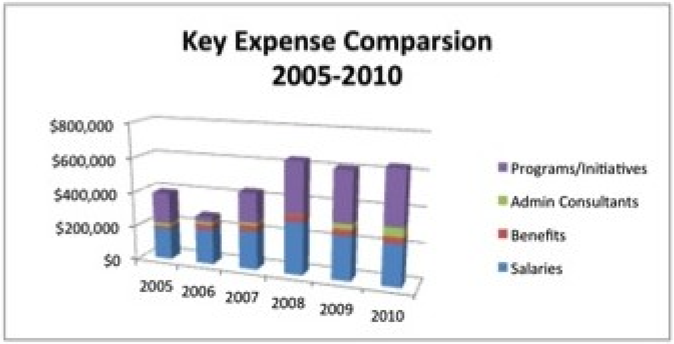
b. List the significant changes needs to be done in this dataset so that it can be used as input to the modelling technique identified in (a)?

c. Show the final dataset structure that can be used as input

Q4. Declutter (remove the noise) the given visualization. Clearly identify the issues and provide the resolutions for them in the textual format. Draw the final revamped visualization, no intermediate visualizations are required. **[6]**

Financial comparisons


Key Income Comparison 2005 - 2010


Key Expense Comparsion 2005-2010

****