SRN | | | | | | | | | | | |

PES

PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

UE15CS322

DECEMBER 2017: END SEMESTER ASSESSMENT

UE15CS322: Data Analytics (SMP SECTION)

Time: 3 Hrs (180 minutes) | Answer All Questions | Max Marks: 100

Instructions: Calculators are permitted.

1) **(20 Marks)**
The table given below describes the percentage contribution to the total revenue generated by five superstores: (refer for question 1 a and 1 b only).

	Total revenue generated (in millions of rupees)	Packed food (%)	Health-care (%)	Cosmetics (%)	Garments (%)	Stationery (%)
Small Bazar	**45**	10	5	35	10	5
Kubhiksha	**10**	25	13	25	NA	5
Hypomart	**5**	30	15	NA	NA	NA
ToughDay	**20**	10	5	30	10	10
Less4More	**20**	10	5	10	NA	10

Whatever revenue is not accounted for in the table above, comes from the sale of Electronic items (i.e., for example, 35% of the revenue of small bazaar comes from the sale of electronic items).

(a) (4 Marks) When analyzing this data, it is recommended that you drop the columns labeled 'Garments' and 'Healthcare'. How would you justify each of these?

(b) (4 Marks) If it is known that 55% of the revenue of Hypomart comes from the sale of Electronic items, how much of the Rs. 50,00,000 revenues generated by Hypomart comes from the sale of cosmetic items? What is the overall percentage contribution of the sale of stationery items to the total revenue generated by the five superstores?

(c) (6 Marks) There are three reputed agencies which are in the business of doing opinion polls and exit polls for elections. The opinion polls are done at different times before the election by asking people their voting intent. Exit polls are done by interviewing people who have actually voted. Based on learning from this course, comment on

(i) How you decide which agency's prediction is more likely to be correct? What statistical parameters are important here? What information you need to know about methodology of polls?

(ii) When three different agencies provide three different results, what technique can you use to give poll of polls result?

SRN | | | | | | | | | | | | |

(d) (6 Marks) Group of students are seeking admission to Medical and Engineering Colleges. The table below gives details on number of students who opted for different subjects in their 12th grade.

Physics (Yes)	Chemistry(Yes)	Maths(Yes)	Biology(Yes)	Total
1000	1000	700	600	1000

Generally, admission to Engineering is given based on actual marks in Physics, Chemistry and Mathematics and admission to Medicine is given based on actual marks in Physics, Chemistry and Biology.
Assuming you have data on marks of all these students, can you come up with a scheme where admission to Engineering and Medicine is available to all students? You may come up with schemes to predict missing marks based on correlation analysis. Illustrate using examples and plots.

Question 2 (20 Marks)

(a) (10 Marks) Here is a data set of test marks and grades for one of the courses. The students in this class could optionally attend tutorials. Apply linear regression model and logistic regression model to predict the chances of a student getting grade A or higher. Comment on the benefits of attending the tutorial based on your analysis.

Test Marks (out of 50)	# of students in the range	Students Attending Tutorials	Students with Grade A or higher who also attended Tutorials	Students with Grade A or higher who did NOT attend tutorial
< 20	10	5	0	0
20-30	40	10	1	0
30-40	25	10	4	1
40-50	25	15	11	7
Total	**100**	**40**	**16**	**8**

(b) (3 Marks) For each scenario below, what type of regression model is most suitable?
   i. Given the data on performance in an aptitude test and CGPA, predict the starting salary of a fresh graduate.
   ii. Given the data on performance in an aptitude test and CGPA, predict whether the student gets a tier-1 job AND an admission to an elite institution for higher studies.

(c) (2 Marks) Give an example where regression model can be used for classification.

(d) (5 Marks) A Post office picks up many letters from different post boxes within a specific zone *Z*. It needs to distribute letters zone *Z*, and forward letters meant for areas outside zone *Z*. In addition, the workload should be divided among postmen such that the overall delivery time is small. Suppose we have software that can accurately scan read the address of each letter. What kinds of clustering algorithms can it use to achieve the given objectives?

SRN | | | | | | | | | | | |

Question 3 (20 Marks)

(a) (6 Marks) The following table gives the user ratings for different items. Find the missing ratings based on the available data.

Items->	M1	M2	M3
User 1	2	?	3
User 2	5	2	?
User 3	3	3	1
User 4	?	2	2

(b) (8 Marks) The table below gives the age of borrower, loan amount and whether the loan was defaulted i.e. not paid back to a bank. A new borrower aged 48 has approached the bank for a loan of USD 142,000. Should the bank lend him? (Hint: You can apply K-NN classification model)

Age	Loan Amount	Default(Y/N)
25	$40,000	N
35	$60,000	N
45	$80,000	N
20	$80,000	N
35	$20,000	N
52	$120,000	N
23	$95,000	Y
40	$62,000	Y
60	$100,000	Y
48	$220,000	Y
33	$150,000	Y

(c) (4 Marks) We have 9 items and 10 transactions. Find support and confidence for each pair that has bread as one of the items. Assume Minimum support to be 2. How many pairs are there?

Transaction Id	Items
10	Bread, Cheese, Newspaper
20	Bread, Cheese, Juice
30	Bread, Milk
40	Cheese, Juice, Milk, Coffee
50	Sugar, Tea, Coffee, Biscuits, Newspaper
60	Sugar, Tea, Coffee, Milk, Juice, Newspaper
70	Bread, Cheese
80	Bread, Cheese, Juice, Coffee
90	Bread, Milk
100	Sugar, Tea, Coffee, Bread, Milk, Juice, Newspaper

(d) (2 Marks) Give an example of classification problem where SVM model is particularly useful.

3/4

SRN | | | | | | | | | | | |

Question 4 (20 Marks)

(a) (8 Marks) With examples, explain ARIMA(p,d,q) model. In which situation, exponential smoothing useful?

(b) (6 Marks) Explain the difference between Auto-correlation Function and Partial Auto-correlation Function with examples. Which function is more useful to study seasonality?

(c) (6 Marks) The table below gives the sales data of Energy Drinks beverage. State your observation by plotting the variation. Comment on the trend for upcoming year.

Quarterly sales of Tiger Sports Drink

	Year							
Quarter	1	2	3	4	5	6	7	8
1	72	77	81	87	94	102	106	115
2	116	123	131	140	147	162	170	177
3	136	146	158	167	177	191	200	218
4	96	101	109	120	128	134	142	149

Question 5 (20 Marks)

We propose to classify all research papers published by PES students and alumni based on the domains to which they belong. The words associated with each domain are taken from the write up on the domain in the PES website. Assume that the full contents of research papers published by PES students and alumni are available for the analysis. The analysis steps are:

Step 1: Tokenize words in each research paper to create a list of all the words in each paper
Step 2: Create a list of words in the write up associated with each domain (the write up associated with a domain is known from the title of the article, which is the domain name)
Step 3: Match the words in the research paper with those appearing in each domain and classify the project as belonging to the domain (one or more) with the most matches.

a) (5 points) Suggest a mechanism to implement Step 1 (i.e., to split a research paper into a list of all words contained in that paper).

b) (5 points) Steps 1 and 2 will yield several words. Suggest a mechanism to eliminate some of the words that are **NOT** indicative of the domain (i.e., a mechanism to extract 'key words' automatically)

c) (5 points) A naïve approach to implement Step 3 would be string matching. That is, for every word in list of keywords extracted from blog posts pertaining to a Unit, compute the number of instances it is found in the project report through character-by-character comparison. Suggest a better strategy to classify each project report.

d) (5 points) Describe strategies (based on the given data) to identify domains that are similar.

4/(,)