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

PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

UE16CS322

DECEMBER 2018: END SEMESTER ASSESSMENT

UE16CS322: Data Analytics(SMP Section)

Time: 3 Hours (180 Minutes) Answer All Questions Max Marks: 100

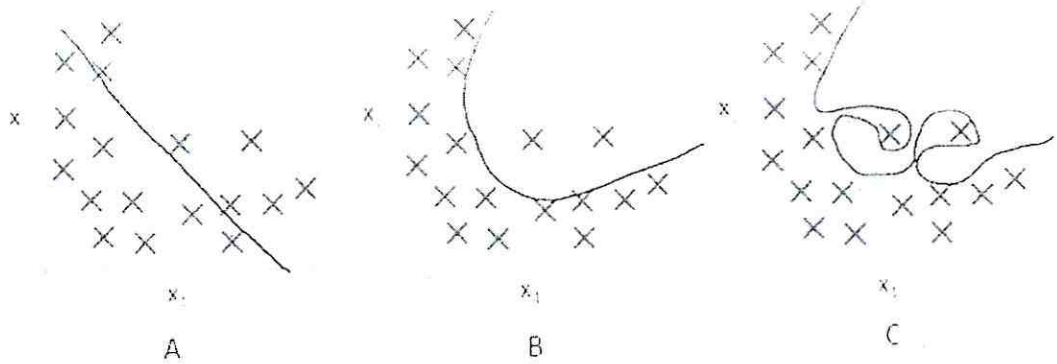
Instructions: Only calculators are permitted. Please show your work-outs of problem clearly.

1	a	The number of customers visiting a shopping mall in Bangalore in the last 20 days is as follows: 232, 277, 261, 173, 283, 197, 251, 212, 213, 213, 229, 164, 219, 196, 186, 247, 244, 269, 216 and 272. Calculate skewness and kurtosis. What do you infer from these? How can the shopping mall use this analysis? (Hint: Kurtosis= $\sum (x_i - \bar{x})^4/n\sigma^4$), Skewness- apply the same for 3rd power)	6
	b	In the above example, the age details of customers were collected and found to be as follows: 0-20 years: 500 20-40: 800 40-60: 500 >60: 200 The mall authorities would like to send 100 emails using clustered sampling and another 100 by stratified sampling to get feedback from people about the Mall. How do you go about constructing samples? What additional information you may need?	4
	c	Explain why with a fixed number of training samples, predictive power decreases with increasing dimensions.	4
	d	Explain Principal Component Analysis taking the example of Iris data-set below. Just explain the intent and outcome expected. Number of Instances: 150 (50 in each of three classes) Number of Attributes: 4 numeric, predictive attributes and the class. Attribute Information: 1. sepal length in cm 2. sepal width in cm 3. petal length in cm 4. petal width in cm 5. class: -- Iris Setosa -- Iris Versicolour -- Iris Virginica	6
2	a	Corruption perception Index and Gini Index (measures inequalities) are given for the following countries.	6

Country	Hongkong	South Korea	China	Italy	Mongolia	Austria	Norway
Corruption Index	77	53	40	47	38	75	85
Gini	53.7	30.2	46.2	32.7	36.5	27.6	23.5

Draw a simple linear regression model and predict Corruption Perception for India with Gini index of 38. Note: Higher the Corruption index better the transparency. Higher the Gini Index means higher inequality. (Hint: b0 = mean of Y – b1*mean of X, b1 =$\sum xy$ – n* mean of x*mean of y/($\sum x^2$ – n *(mean of x)^2)

SRN

	b	Below are three scatter plots (A, B, C) and hand-drawn decision boundaries for logistic regression. Answer the questions that follow:  A B C i. Which figure is over-fitting the training data the most and why? ii. In which model is the training error maximum and why? iii. Which model is more robust than the other two models and why?	6
	c	Before you apply regression model to a data-set what assumptions do you need to validate? What kind of data-cleaning can one do? Explain	4
	d	Explain the following: (i) Difference between regression and correlation (ii) It was found that married men earn more money by analyzing data using regression model. Can we infer that one should get married to earn money?	4
3	a	You are given the transaction data shown in the Table below from a fast food restaurant. There are 9 distinct transactions (order:1 – order:9) and each transaction involves 2-4 meal items. There are a total of 5 meal items that are involved in the transactions. Meal items are names M1-M5 for simplicity	8

Meal Item	List of Items
Order 1	M1, M2, M5
Order 2	M2, M4
Order 3	M2, M3
Order 4	M1, M2, M4
Order 5	M1, M3
Order 6	M2, M3
Order 7	M1, M3
Order 8	M1, M2, M3, M5
Order 9	M1, M2, M3

Assume minimum support is 2/9 and the minimum confidence is 7/9.
i. Identify all the frequent item sets involving M5 as one of the meal items. Calculate the support for each item.
ii. What is the support and confidence of the rules and identify which rule is stronger: M1, M5 -> M2 Or M2, M5 -> M1

	b	Consider the 7 observations in a two-dimensional space. {(2,2), (4,4), (6,6), (1,4), (4,0), (5,5), (9,9)}. If we run the k-means clustering algorithm, to divide these data-point into 3 clusters. What would be the clusters after the first iteration for each of the data- point. Use Manhattan distance to calculate the distance between clusters. Use (4,4), (1,4) and (5,5) as initial seeds. What would the centroids for the second iteration?	6

SRN

	c	Explain with an example how do you choose nodes to do a split in decision trees. What measures can you use?	6
4	a	With Example, explain ARIMA(p,d,q) modeling.	6
	b	How do you detect the different time series components (Trend, Cyclic, Seasonal and Irregular components) in time series data? Explain with a rough sketch of the corresponding graphs	6
	c	Before conducting time series analysis, we need to consider whether the variables are stationary. What is stationarity? How do you determine if your data is stationary or not? How do you deal with non-stationary series?	4
	d	Explain exponential smoothing with an example.	4
5	a	You have written a research paper and looking for a suitable journal to publish. The only information you have is already published papers in these journals. How do you go about choosing the journal?	6
	b	Explain with examples difference between synonymy and polysemy	4
	c	Write an algorithm to help break paragraph into sentences. Test it for the following sentence. Yes, Mr. Anurag. You need a D.L. to drive.	6
	d	What approach can you use to classify customer feedback as positive and negative?	4