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

PES UNIVERSITY

PES University, Bengaluru
(Established under Karnataka Act No. 16 of 2013)

UE17CS322

DECEMBER 2020: END SEMESTER ASSESSMENT (ESA)
B TECH 5TH_ SEMESTER

UE17CS322 – Data Analytics

Time: 3 Hrs	Answer All Questions	Max Marks: 100

1	a)	Use a flowchart to summarize the following procedures for attribute subset selection: (a) stepwise forward selection (b) stepwise backward elimination (c) a combination of forward selection and backward elimination	6
	b)	What are the value ranges of the following normalization methods? (a) min-max normalization (b) z-score normalization (c) z-score normalization using the mean absolute deviation instead of standard deviation (d) normalization by decimal scaling	4
	c)	Consider the following data (in increasing order) for the attribute age: 13, 15,16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46,52, 70. Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps. Discuss on the effect of this technique for the given data.	6
	d)	How might you determine outliers in the data?	4
2	a)	Professor Bell at Bellandur University, Bangalore believes that the cumulative grade point average (CGPA) of the students is negatively correlated with usage (measured in average minutes per day) of smart phones. Table 1 shows the CGPA and smart phone usage in minutes per day of 40 students. (a) Calculate the Pearson correlation coefficient between CGPA and mobile phone usage of students. (b) Conduct a hypothesis test at a = 0.01 to check whether CGPA and mobile phone usage are negatively correlated. (c) Professor Bell believes that the correlation is less than −0.4. Conduct a hypothesis test at a = 0.1 to check whether the claim is correct. Table.1: Data of CGPA and mobile phone usage (Average minutes per day)	6

CGPA	2.65	2.25	1.86	1.47	2.10	1.94	2.71	1.83	2.65	2.04
Phone Usage	75	89	65	136	95	103	74	109	7	98
CGPA	2.54	2.16	2.28	2.47	2.18	2.57	1.97	2.87	2.10	3.28
Phone Usage	60	93	88	81	92	78	102	70	95	89
CGPA	2.78	2.44	2.50	2.24	2.01	2.17	2.20	2.05	1.63	1.87
Phone Usage	72	82	107	80	89	100	92	91	98	123
CGPA	2.28	2.63	2.86	2.24	2.44	2.69	2.22	3.07	1.77	3.03
Phone Usage	88	76	70	89	82	74	90	65	113	66

SRN

	b)	A regression model is developed between corruption perception index and per capita income (in US dollars) based on data on 20 countries. Regression model output obtained through Microsoft Excel is shown in Table 2. Note that Table 2 shows only partial output of the model developed. TABLE 14. Regression between corruption perception index (Y) and per capita (X) Table 2. Corruption Index and Gini Index.	4

SUMMARY OUTPUT						
Regression Statistics						
Multiple *R*						
R Square						
Adjusted *R* Square						
Standard Error	10.94929					
Observations	20					

ANOVA						
	df	*SS*	*MS*	*F*	Significance *F*	
Regression	1	5918.236				
Residual	18	2157.964				
Total						
	Coefficients	Standard Error	*t*-Stat	*p*-value	Lower 95%	Upper 95%
Intercept		6.496415			5.773095	33.07002
Per Capita		0.00016			0.000788	0.001461

(a) What proportion of the corruption perception index is explained by per capita?
(b) What is change in the value of corruption perception index for every one-dollar increase in per capita?
(c) Is there a statistically significant relationship between corruption perception index and per capita at a = 0.01?
(d) What is the average corruption perception index when per capita is $30,000. What is the corresponding 95% confidence interval?
(e) Per capita of a country is $30,000. What is the probability that the corruption perception index of this country is less than 50?
(f) Which of the following statements are true based on the model shown in Table 13?
(i) Corruption perception index and per capita are positively correlated.
(ii) Corruption perception index and per capita are negatively correlated.
(iii) There is no correlation between corruption perception index and per capita.

	c)	Define each of the following data mining functionalities: characterization, discrimination, association and correlation analysis, classification, regression, clustering, and outlier analysis. Give examples of each data mining functionality, using a real-life database that you are familiar with.	6
	d)	Explain the difference and similarity between discrimination and classification, between characterization and clustering, and between classification and regression.	4
3	a)	Briefly explain the different components of a Time-Series Data	6
	b)	How do we Identify Seasonality?	4
	c)	Quarterly demand for certain parts manufactured by Jack and Jill company is shown in following Table	6 (3+3)

Quarterly demand

Year	Quarter	Value
2012	Q1	75
	Q2	60
	Q3	54
	Q4	59
2013	Q1	86
	Q2	65
	Q3	63
	Q4	80
2014	Q1	90
	Q2	72
	Q3	66
	Q4	85
2015	Q1	100
	Q2	78
	Q3	72
	Q4	93

(a) Calculate the seasonality index for different quarters using the first 3 years of data.
(b) Develop forecasting models using moving average, single exponential smoothing, and an appropriate ARMA model after de-seasonalizing the data (assume multiplicative model, $Yt = Tt \times St$).

	d)	Forecast the demand for 2015 (all four quarters) using moving average, exponential smoothing, and ARMA.(From Question 3.c) Calculate RMSE, MAPE.	4

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

4	a)	What is a recommender system? In what ways does it differ from a customer or product-based clustering system? How does it differ from a typical classification or predictive modeling system?	5
	b)	Describe one method of collaborative filtering. Discuss why it works and what its limitations are in practice.	5
	c)	The movie ratings given by 4 customers (C1, C2, C3, and C4) on five movies (A, B, C, D and E) are given in Table 14.17.	4
	d)	Customer feedbacks on 5 training programs (on a 5-point scale) by 6 customers are provided in the following Table	6

TABLE 14.17 Movie ratings by customers

Movies → Customer↓	A	B	C	D	E
C_1	4	1	3	2	4
C_2	3	2	4	4	2
C_3	3	3	4	4	4
C_4	3	3	2	3	3

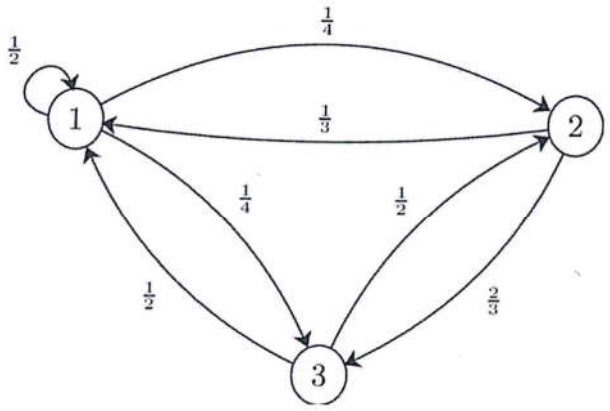
Use cosine similarity to find among customers C1, C2, and C3, who is the closest to customer

Tables5. Feedback on training programs

	M_1	M_2	M_3	M_4	M_5
C_1	2	4	2	4	3
C_2	4	3	2	4	5
C_3	1	2	3	2	4
C_4	4	4	2	4	3
C_5	2	1	2	2	3
C_6	2	1	1	4	4

(a) Use cosine similarity to identify the customer who is closest to customer 1.
(b) Calculate correlation between different customers. Which customer has the highest correlation with customer 1?
(c) What is your conclusion based answers to questions (a) and (b)?

SRN

5	a)	Consider the markov chain as shown below,  Give answers for the following questions: a. This Markov chain irreducible? b. Is this Markov chain aperiodic? c. Find the stationary distribution for this Markov chain. d. Is the stationary distribution a limiting distribution for the Markov chain?	6
	b)	The number of customers arriving at a departmental store can be modelled by a Poisson process with λ=10 customers per hour. Find the probability that there are 2 customers between 11:00 and 11:20 am	4
	c)	i). When a Markov chain is said to be ergodic? ii).Define a confounding variable?	4 (2+2)
	d)	How to reduce Confounding Variables? And What Conditions Cause Omitted Variable Bias?	6 (3+3)