PES UNIVERSITY

PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

UE18CS312

OCTOBER 2020: IN SEMESTER ASSESSMENT B Tech FIFTH SEMESTER TEST – 1

UE18CS312 (4 credit subject) - Data Analytics

Time: 2 Hrs	Answer All Questions	Max Marks: 60

1. a) An online certification course has been offered to students in the fifth and seventh semesters of CSE. The number of registrations and number of successful certifications across the country at the end of each month as recorded by the course is provided below: **4 (2+2)**

Month	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
Number of registrations	44	101	386	4,904	12,106	74,696	1,02,458	12,524
Successful certifications	6	59	174	359	18,036	72,599	96,239	6,980

(i) If we use a Cox-comb plot to visualize this data, how many sectors would this plot have and how would we represent the data provided in the table?

(ii) A potential registrant wants to answer the question: "Is the increase in the number of certifications between the fifth and seventh semester students statistically significant?" Assuming the detailed data for fifth and seventh semesters is available, outline the steps of the approach one might take to answer this question.

b) In the data shown in Question 1(a) above, **3**

(i) Is there any anomaly? Substantiate your answer with a reason.

(ii) The organizers of the course have realized that data has not been recorded during weekends in the month of April. Suggest a method to fill in the missing values.

c) The range of scores of students for various components of their project submission is (3,8). If the marks scored are rescaled to a new range, (16, 25), what will a score of 7 in the (3,8) scale map to in the new range? **3**

2. a) Briefly explain the sampling technique(s) used in each of the following cases: **4 (2+2)**

(i) A Kitkat factory produces ten different flavours of the chocolates and has twenty assembly lines (two for each flavour). A taste tester selects a random chocolate bar from every other line.

(ii) A restaurant has placed a feedback card on every table and allows diners to choose whether they would like to provide a feedback on behalf of their group or not.

b) For the following examples, identify the datatypes as numeric/ categorical, ordinal/ interval/ ratio and discrete/ continuous (as applicable) **3**

(i) Movie rating on a scale of 1 to 5

(ii) When booking a flight ticket, response to whether the wheelchair service would be required for the passenger (yes/ no)

(iii) Temperature (recorded in Centigrade from various regions around the world)

c) Twenty engineers and twenty pilots were subject to tests and scores were measured for the following six features: (i) Intelligence (ii) Conformance to procedure (iii) Eyesight (iv) Hearing (v) Sensory motor coordination and (vi) Perseverance. **3 (2+1)**

Briefly outline the steps to extract two principal components from this data to visualize the two

groups of twenty points in the 2-dimensional rectangular plane.

3. a) Taste testers Aman and Mani have rated the quality of food at a restaurant on six days in the week as follows: — 4 (2+1 +1)

Day	M	Tu	W	Th	F	Sa
Rating(Aman)	4	2	3	5	1	3
Rating(Mani)	3	3	2	5	2	2

Given: mean and standard deviation of ratings: $\mu_{Aman} = 3$, $\sigma_{Aman} = 1.291$, $\mu_{Mani} = 2.833$, $\sigma_{Mani} = 1.067$, correlation coefficient(Aman, Mani) = 0.7258

(i) What are β_0 and β_1 if we must predict Aman's rating in terms of Mani's rating using the simple linear regression with the following model?

$$\textbf{Rating(Aman)} = \beta_0 + \beta_1 \cdot \textbf{Rating(Mani)}$$

(ii) What is the coefficient of determination for this model?

(iii) How can we measure the influence that Mani's rating of the food on Thursday has on the model? (Suggest the test or statistic that can be used for this.)

b) The correlation between two variables (#views for a video and average #videos posted per month) on a video sharing platform is found to be positively correlated. Answer the following questions and (briefly) substantiate your answer: — 3

(i) Is it necessarily true that the Pearson's correlation coefficient between #postings/ month and #views on a video would to be closer to 1 than it is to 0 for this data?

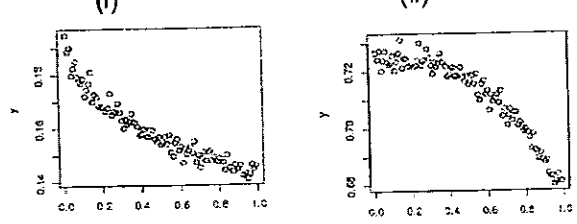
(ii) Can we assume there is no cause-effect relationship between #postings per month and #views on a channel because correlation does not imply causation?

c) For each of the following scatterplots, state whether the data is suitable for linear regression — 3

$$y = \beta_0 + \beta_1 \cdot x$$

and, if it is not, what transformation(s) may be applied to the variable(s) make this data amenable for modeling with linear regression.

(i) (ii)



4. a) Write the linear algebraic equation for computing an estimate of the Beta vector in a multiple linear regression system to predict 4 dependent variables using 5 independent variables. In the table given below, identify the features that are significant (for an alpha = 0.01). If there is insufficient data to do this, list out what other data is necessary to determine the significance of regression coefficients. — 4 (2+2)

Term	Coef	SE Coef	T	P
Constant	389.166	66.0937	5.8881	0.000
X_1	2.125	1.2145	1.7495	0.092
X_2	5.318	0.9629	5.5232	0.000
X_3	4.22	0.3	14.06	0.043
X_4	-24.132	1.8685	-12.9153	0.000
X_5	-17.201	1.333	-12.9039	0.004

4.	b)	Rajesh has designed a logistic regression classifier to predict the likelihood of stars being visible in the night sky based on the humidity reported on any day: logit (p) = **log(p/(1-p)) = β_0 + β_1*humidity**, where p is the probability stars are visible at night. Given that β_0 = 1.8185 and β_1 = -0.0665, answer the following questions: (i) What does the value of β_0 mean? (ii) If humidity on a day = 25, what is the probability with which stars are visible in the night sky according to this model?	3 (1+2)
	c)	In a collection of 1000 small rocks collected on a river bed, 100 happen to be precious stones. All the 100 precious stones along with 100 other rocks have been classified as 'precious stones' by a logistic regression model. Write the entries of the confusion matrix for this classifier, clearly labeling the rows and columns. What further steps should be taken to plot the receiver operator characteristics (RoC) for this logistic regression model?	3 (2+1)
5.	a)	With a schematic sketch, briefly describe the key characteristics of the level, trend and seasonality components of an additive time series data. What are cyclic components and, why are they usually not accounted for in models for time series data?	4 (3+1)
	b)	For the data given below, use MAPE to compare the forecast accuracy of single exponential smoothing (SES) with alpha = 0.7 with the forecast accuracy of the simple moving average (SMA) with a window size = 3 for time points t=5,6,7. [You can use the values of y that are available to make the forecasts for SMA and for SES assume the forecast, $F_4=y_4$.]	3
	c)	Suggest an application for each of the following techniques to model time series data (i) Croston's method (ii) Holt-Winter's method (iii) ARIMA	3
6	a)	Write the equation corresponding to the two models given below (explain the symbols clearly): (i) ARIMA(0,1,0) (ii) ARIMA(1,0,1)	4 (2+2)
	b)	Which model is better and why? (3+3)	6 (3*2)

T	1	2	3	4	5	6	7
y_t	10	11	12	16	17	19	20

	Statistic	Model A	Model B	Better: Model A or Model B?	Why?
1	AIC	258.24	251.42		
2	R^2	0.98	0.91		
3	RMSE	0.048	0.051		