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

PES UNIVERSITY

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

UE18CS312

DECEMBER 2020: END SEMESTER ASSESSMENT (ESA)
B.TECH FIFTH SEMESTER

UE18CS312 – Data Analytics (4 credit Elective course)

Time: 3 Hrs	Answer All Questions	Max Marks: 100

1 a) What factors help increase attendance in the primary schools in villages? The Education Ministry has asked you to help them design a survey to gather this information. **(6 marks, 2+4)**

(i) Who would you recommend collecting the data from? (suggest any two sources or demographics that could provide a meaningful insight)

(ii) Suggest any two variables (or attributes) that you would collect and also identify the type of data (categorical/ numeric and ordinal/ interval or ratio) for the two features.

b) The following table contains the gross collection and budget (both variables are reported in crores of rupees) for ten movies released in 2020 and declared a 'Success' in the box office: **(4 marks, 2+2)**

Collection	60	50	35	50	65	270.15	197.3	39.3	193.2
Budget	40	40	25	38	45	125	105	236	100

(i) Suggest a visualization that will help us easily infer the 'profit' earned by the movie

(ii) It is postulated that higher the budget, higher the box office collection. What visualization technique (or graph) will help us easily infer the relationship between collection and budget? Briefly explain.

c) Given an example for each of the following types of error in data entry and suggest how each can be handled: **(4 marks)**

(i) Incomplete data (ii) inconsistent data

d) Data gathered from a batch of students is recorded below: **(6 marks, 4+2)**

	Knows of Dragon Ball Z	**Does not know of Dragon Ball Z**
Knows of Pokémon	250	200
Does not know of Pokémon	50	1000

(i) What is the chi-square (χ^2) statistic for this data?

(ii) How can this be used to determine whether knowing of Pokémon is independent of knowing of Dragon Ball-Z?
(You may briefly describe the steps to be followed for this.)

Note: You may use the following facts in your computation:

$$\chi^2 = \sum_{i,j} \frac{(observed_{i,j} - expected_{i,j})^2}{expected_{i,j}}$$

$$expected_{i,j} = \frac{\sum A=a_i * \sum B=b_j}{N}$$

where A is the first variable (i.e., the two rows)
and B is the second variable (i.e., the two columns)
and N is the total number of samples in the study

SRN

2	a)	The following table shows the result of multiple linear regression used to predict the number of minutes students are willing to participate in an after-school activity based on multiple factors. Answer the following questions:		6 (3*2)

	Beta	Std error	Standardized Beta	t	p
Constant	17.23	50.16		0.34	0.73
Challenge	-5.88	8.88	-0.06	-0.66	0.51
Instant enjoyment	-20.33	18.13	-0.14	-1.12	0.02
Novelty	9.95	10.78	0.09	0.92	0.001
Exploration	28.33	11.00	0.25	2.58	0.03

(i) Write an equation to represent this model

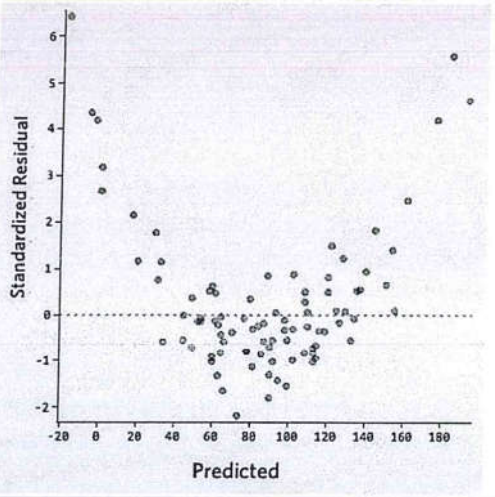
(ii) If the significance value alpha = 0.5, what are the features that would be dropped?

(iii) In the simple linear regression model below, what does the error term, ε represent?

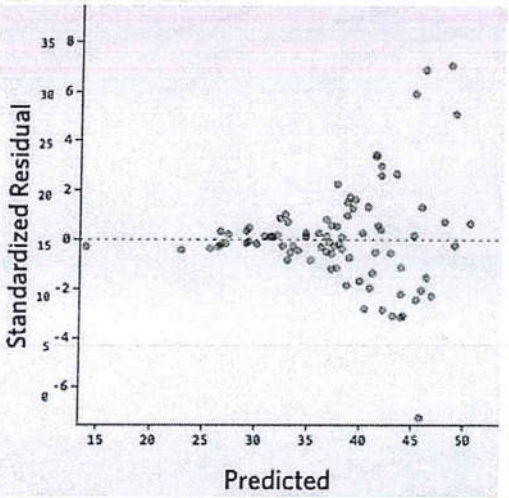
Novelty = 16.43 + 5.22*Exploration + ε

	b)	In a sample of 100 engineers, 90 stay up all night the day before an exam whereas in a sample of 80 doctors, only 20 stay up all night the day before an exam. (i) What is the odds ratio of an engineer staying up all night the day before an exam to a doctor staying up all night? (ii) Suggest any one approach to deal with categorical variables when designing a regression model. (For example, consider a month's worth of data from the postal department on zip code (the explanatory variable) and the number of letters delivered to that zip code in a day (as the predicted variable). How would we use zip code in building a regression model for this data?)	4 (2+2)
	c)	Based on these residual plots from two liner regression models, can the two models be considered good? If there is any problem you can diagnose based on the plots, briefly explain what that is. (The dotted line indicates the zero reading on the vertical axis in both plots.)	4 (2+2)

Residual Plot 1



Residual Plot 2



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

	d)	A logistic regression model to categorize actions in a video clip yielded the following confusion matrix. Answer the questions below: **Predicted class** **Actual class** |	**Dance**	**Martial arts**	**Pilates**
Dance	40	30	10
Martial arts	20	50	10
Pilates	10	10	60	 (i) What is the recall (or true positive rate) for Martial arts? (ii) What is the precision for Pilates? (Of all the clips labeled as Pilates, how many were truly clips of Pilates?) (iii) What is the average accuracy of this classifier?	6 (3*2)
3	a)	Briefly answer the questions: (i) With a schematic sketch of a stationary signal, list the characteristics of a typical stationary signal. How can we test whether a given time series is stationary? (ii) What are the typical values of the parameter d (for differencing) in an ARIMA model and how can we determine the other two parameters p (order of the autoregressive component) and q (order of the moving average component) for an ARIMA model? Explain briefly.	8 (4+4)
	b)	What model would be most appropriate in each of the following scenarios and why? (i) Predicting the yield of a seasonal crop (ii) Predicting the purchase of school supplies (iii) Predicting the stock price of Netflix	6 (2*3)
	c)	What do the following statistics tell us about a time series model? Answer each of the questions below: **(i) Theil's coefficient – what is this statistic used for and what does it mean if U=1?** $U = \frac{\sum_{t=1}^{n}(Y_{t+1} - F_{t+1})^2}{\sum_{t=1}^{n}(Y_{t+1} - Y_t)^2}$, where Y is the actual value, F is the forecast and t is the time index **(ii) Ljung-Box test – what is this test used for and why do we need p and q, the AR and MA lags of the ARIMA model for this test?** $Q(m) = n(n+2)\sum_{k=1}^{m}\frac{\rho^2}{n-k}$, where n is the number of observations in the time series, m is the total number of lags and ρ the autocorrelation of lag k **(iii) BIC – what is this used for and how is it different from AIC?** $BIC = -2LL + Kln(n)$, where LL is the log likelihood function, K is the number of parameters estimated and n is the number of observations in the sample	6 (2*3)

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

4	a)	What is the most appropriate recommendation system for each of the following applications? Explain in a line why this is most appropriate. (You may choose from: content-based, collaborative filtering, knowledge-based (case based, query based or constraint-based) and apriori based association rule-mining) (i) PlanMyTrip – a vacation planning website (ii) Facebook friend recommendations to a new user (iii) Recipes on a cooking site (iv) Predicting whether a particular poll will be answered by a customer who has been active on the news site for the past eight months.	8 (2*4)
	b)	In the context of designing a recommendation system, answer the following questions: (i) In what way is entropy a better criterion for splitting a node than either Gini index or classification error? (ii) If the confidence of an association rule is high, can we assume the rule would be interesting? Substantiate your answer briefly.	6 (3+3)
	c)	The Department of CSE would like to form review panels for the Capstone projects. Each team has submitted a title of their project and a brief description of the problem statement. (i) Outline the steps to cluster similar problem statements together. (ii) There may be multiple configurations of clusters that come up from step (i) above. Suggest an evaluation criterion that will ensure we select a configuration that has maximum inter-cluster distance and minimum intra-cluster distance. Briefly explain how the criterion you suggest can help test these conditions (of minimum intra-cluster distance and maximum inter-cluster distance).	6 (4+2)
5	a)	(i) What is the canonical form of an absorbing state Markov Chain? State what each component matrix stand for. (ii) The number of flights cancelled by an airline is modelled using a Markov chain. The state transition matrix between flight cancellations is shown below: (0 – no cancellations, 1-3 represents the number of cancellations on a day):	8 (4+4)

	0	1	2	3
0	0.45	0.30	0.20	0.05
1	0.15	0.60	0.15	0.10
2	0.10	0.30	0.40	0.20
3	0	0.10	0.70	0.20

If there are no cancellations initially, what is the probability there will be at least one cancellation after two days?

	b)	Check whether the matrices below are regular. (Note: A regular matrix is one that has all positive entries for some power of the matrix; it is not required to check beyond a power of 4.)	6 (3+3)

(i)

0	0.5
1	0.5

(ii)

0	1
1	0

	c)	(i) Suggest a method to check whether a hidden variable is confounding. (ii) Microsoft Stream is considering introducing a 'Download all' button for all video clips in a class group. The advantage is that students do not have to click on each link to download the recorded session for each class. But the disadvantage is that the file is extremely large and students may not have the bandwidth to download the material with one click. As an expert on Data Analytics, outline the steps you would take to design this test for the MS Stream team.	6 (3 +3)