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

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

UE18/19CS312

DECEMBER 2021: END SEMESTER ASSESSMENT (ESA)
B.TECH FIFTH SEMESTER

UE18/19CS312 – Data Analytics

Time: 3 Hrs	Answer All Questions	Max Marks: 100

1 a) What factors help increase innovation and entrepreneurship in youngsters? The Ministry of Education 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 answer to the question)

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

b) The following table contains the type of activity and average number of steps in 1 minute as measured by PESFit, a fitness band developed by PES worn on one hand and a commercial fitness band (referred to as FitBand) that is worn on the other hand by the same person throughout each of the activities: **(8 marks, 2+2+2+2)**

Activity	Swim-ming	Basketball	Tennis	Skipping (2 jumps/ sec.)	Soccer	House cleaning	Golf	Yoga
PESFit	165	140	155	114	145	120	130	95
FitBand	171	164	242	NA	145	126	134	72

(i) Suggest a visual representation for this data to help make quick inferences

(ii) A gym instructor feels the intensity of activity and hence, the calories burned, can be inferred from the number of steps. Give one reason to convince the gym instructor this is not true.

(iii) The commercial Fitness Band used in this study ran out of charge just before skipping and hence, no reading has been recorded for this activity. Is this considered MAR, MCAR or MNAR? Substantiate your answer with a reason.

(iv) What is the best way to deal with the missing data in this table?

c) The marks of twelve students is as follows: 6, 10, 11, 11, 12, 12, 13, 13, 13, 14, 14, 15. The average is found to be: 12. **(6 marks, 2+2+2)**

(i) Is this data skewed? Substantiate your answer with a reason.

(ii) How does the average change if 6 is a typographical error and that entry was actually 9?

(iii) If the data from 9-15 should be mapped to a new range 25-40 (both included), what is 13 in the new range?

2 a) List any three assumptions that are deemed necessary for multiple linear regression and briefly explain how you would check whether this assumption holds for a given dataset. **(6 marks, 3*2)**

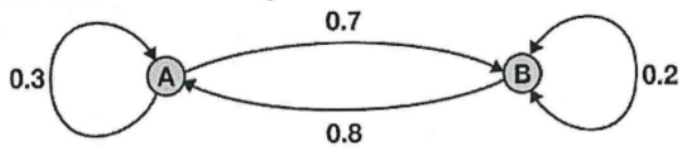
SRN | | | | | | | | | | | |

	b)	A logistic regression model for the probability of becoming a successful entrepreneur if someone in the family is also an entrepreneur has the intercept **-1.93** and coefficient of **0.38** for x, the binary independent variable (x = 1 for at least one family member is an entrepreneur and x=0 for no family member is an entrepreneur): (i) What does the intercept = -1.93 mean? Interpret this value, with a brief explanation. (ii) What is the change in the odds ratio that associates a family member being an entrepreneur to a person becoming a successful entrepreneur? (You may use the fact: ln(odds ratio) = $\beta_0+\beta_1 x$, where x is the independent variable.) (iii) In a dataset of 100 people, the model predicts 94 can become successful entrepreneurs of which 9 are found to be false positives and the number of correct predictions for 'cannot become a successful entrepreneur' = 1, what is the F1-score of this model?	8 (2+3+3)
	c)	In the context of regression models, answer the following questions: (i) Does the L1 regularization in Lasso regression help in feature selection? Explain. (ii) If the Pearson's correlation between two variables = 0.20, can we conclude that the two variables are not at all related to each other? Briefly explain.	6 (3*2)
3	a)	Briefly answer the questions: (i) Sketch the figure of a stationary signal and one of a non-stationary signal, identifying the features that make the signal stationary and nonstationary, respectively. (ii) Tabulate the guidelines for selecting model parameters for AR(p) and MA(q) models using the autocorrelation function (ACF) and the partial autocorrelation function (PACF)	8 (4+4)
	b)	Suppose the attendance at Hogwarts School during the months of November and December 2021 are as follows: 150, 200. (i) What is the attendance predicted to be in January 2022 based on single exponential smoothing for this data, with alpha = 0.4? (You may assume the value of the forecast for November, F_0= actual attendance Y_0 in November) (ii) Professor Marazion wants to order new cauldrons for the class on potions and has asked you to build a model to predict the requirement. Which model would you opt for and why? (iii) The seasonality index (SI) for receiving a Howler (a letter of reprimand) from home in October after midterms is found to be 1.37. What does SI=1.37 mean?	6 (2*3)
	c)	Write the equation form of the following models: (i) ARIMA(2,0,0) and (ii) ARIMA(0,1,0)	6 (2*3)
4	a)	The table details transaction data on the purchase of Plush toys from a Pokemon themed store. Given that any itemset with the support count >= 2 is considered frequent, answer the following questions: (i) Which items are not frequent 1-itemsets? (ii) What are the frequent 2-itemsets formed? (iii) Compute the confidence for the association rule {Eevee, Mew}→{Pikachu}. Is this symmetric? (iv) What does Confidence = 1 for an association rule mean? (Note: Support = support count/ number of transactions, Confidence (A→B) = support (A union B)/ support(A))	8 (2*4)

TID	Items bought
T1	Charizard, Onix, Mew
T2	Pikachu, Eevee, Mew
T3	Onix, Pikachu, Eevee, Mew
T4	Pikachu, Eevee
T5	Eelektrik

SRN

	b)	The table below shows the rating of three movies (m1, m2 and m3) by four users (u1,u2, u3 and u4). (i) What is the item-item cosine similarity (CS): CS(m1, m2) and CS(m2,m3)? (ii) What is the rating for movie m2 by user u1 using item-item similarity? (iii) What is the cold start problem in collaborative filtering? How can this be remedied? (Suggest any one approach.) (You may use the fact CS(a, b) = transpose(a)*b/ (sqrt(transpose(a)*a*transpose(b)*b))	6 (2+2+2)
	c)	(i) How can DBSCAN be used to identify 'noise' points or outliers in the given data? (ii) Why is 'novelty' important for a recommender system? How can this be evaluated?	6 (3+3)
5	a)	The state transition diagram of a 2-state Markov Chain is shown below. (i) What is the transition probability matrix, **P**, corresponding to this diagram? (ii) Find the stationary distribution $[\pi_1 \ \pi_2]$ for this matrix (Hint: $\pi_j = \sum_{k=1}^{m} \pi_j P_{kj}$ and $\sum_{k=1}^{m} \pi_k = 1$) (iii) Does the adjacent matrix have a stationary distribution? Substantiate your answer with a reason.	8 (2+3 +3)
	b)	Answer the following questions (i) Provide any two suggestions to ensure the analysis of an A/B test is well done. (ii) Suggest any two methods to infer hidden variables in a study.	6 (3+3)
	c)	The number of flights cancelled by an airline daily is modelled using a Markov chain. The transition probability matrix for the states 0, 1, 2 and 3 cancellations per day is shown below Revenue loss (in lakhs of rupees) due to cancellations is shown below: (i) What is the initial state vector if there are no cancellations initially? (ii) What is the probability there will be at least one cancellation the next day?	6 (2+4)



	m1	m2	m3
u1	2	?	3
u2	5	2	NA
u3	3	3	1
u4	NA	2	2

0	1
1	0

	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

State	0	1	2	3
Loss	0	4.5	10.0	16.0